2016

Covariate Adjustment in Non-Inferiority Trials: Implications for Type I Error

Katherine Nicholas
Medical University of South Carolina

Covariate Adjustment in Non-Inferiority Trials: Implications for Type I Error

Katherine Nicholas

A dissertation submitted to the faculty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies.

Department of Public Health Sciences

2016

Approved By:

Valerie Durkalski

Chair, Advisory Committee

Wenle Zhao

Sharon Yeatts

Jody Ciolino

Keith Borg

Patrick Mauldin

Acknowledgements

I would like to thank all of the wonderful people who have made this work possible including my friends and family who have stuck with me through all of the trials and tribulations of graduate school, my major advisor and dissertation committee for sharing their time and their ideas, Ms. June Watson for always keeping me on track throughout the process, and my fellow students for sharing their strengths and their support. I would also like to acknowledge the support of my research funding via the Neurological Emergencies Treatment Trial Network (U01NS059041) funded by the National Institute of Neurological Disorders and Stroke (NINDS) without which none of this would have been possible.

Table of Contents

Lists of Tables

Lists of Figures

Abstract

There has been little work to date regarding the proper use of covariate information in non-inferiority trials. Too often knowledge obtained in the superiority setting is applied directly to the non-inferiority setting. However, due to the reversal of the hypotheses and the consequent reversal of the implication of error probabilities, this is a dangerous practice. The current work demonstrates that in both superiority and non-inferiority, failure to adjust for important covariates results in estimates of treatment effect that are biased towards zero with standard errors that are deflated. However, as no treatment difference is approached under the null hypothesis in superiority and under the alternative in non-inferiority, this results in decreased power and nominal or conservative (deflated) type I error in the context of superiority, but inflated power and type I error under non-inferiority. This occurs regardless of adjustment at randomization.

Generally, it is advised that covariates requiring adjustment be specified before the start of the trial. However, important prognostic factors are not always

known in advance. Thus, a joint statistic for the identification of important covariates based on the simultaneous assessment of influence on outcome and disparity across treatment groups is developed for the non-inferiority setting. This statistic, when calculated for all available covariates in a trial, can be used to rank them according to importance. This ranking can be used to identify the subset that will optimize the tradeoff between the change in the point estimate of the treatment effect and its precision while preserving type I error. This method is applied to the Rapid Anticonvulsant Medication Prior to Arrival Trial (RAMPART) and its pediatric cohort.

Chapter 1

Introduction

With the increase in the number of efficacious therapies available, the non-inferiority trial is growing progressively more popular. It is often used when a current standard of care has already been shown to be effective, but an alternative many offer some added benefit such as lower cost, greater ease of administration, or greater tolerability. Thus, a non-inferiority trial seeks to address the question of whether a new drug or intervention is "no worse than" a current standard of care by some predetermined margin, the non-inferiority margin. Under this design, the null hypothesis states that the new drug is worse than the standard and the alternative states that it is nearly equal by some pre-defined amount.[5] These hypotheses are often formally tested using a confidence interval approach in which the null is rejected if the lower bound crosses the pre-determined non-inferiority margin.[13,21] Due to this reversal in the hypotheses, the implications of type I and type II error are also reversed.[5] An FDA guidance[41] exists for non-inferiority trials, which outlines several key design issues including

the proper specification of hypotheses, the choice of active control, the determination of the non-inferiority margin, and the appropriate analysis method. However, gaps remain in this guidance as well as in the statistical and clinical literature, including recommendations for the proper use of covariate information.

In a PubMed search of phase-III clinical trials published in the last 10 years containing the word "non-inferiority", 44.8% of 58 trials utilized complete randomization and unadjusted primary analyses. Covariate adjusted randomization and appropriate analysis of covariance or stratification was used in 20.7% of trials. Covariate adjustment at randomization, but not at analysis was done in 25.7%, and adjustment at analysis but not at randomization was done in 8.6% of the trials. This review of current practice shows that there is still considerable variability in the way that clinical trialists use covariate information, a finding that is not surprising considering the lack of attention it has received in the non-inferiority literature. Some feel that primary analyses should always be unadjusted as this offers greater interpretability. Others believe that tests of baseline imbalance can be used to determine which covariates require adjustment, while still others argue that all known prognostic covariates should be included at both randomization and analysis.[1,31,32,35,36]

An issue arises, however, when important prognostic variables are not known in advance. In this case, a data-driven method for choosing covariates based on both disparity with treatment allocation and influence on outcome is of

value. This can be accomplished through a joint statistic that combines these two factors to maximize the tradeoff between the change in the point estimate itself and its precision.[3,6,7] While such data-driven techniques could have important implications for type I error, the statistic can be designed to reduce this potential such that covariates with disparity and influence in the same direction are favored. Thus, the treatment estimate is always conservative. In practice, this statistic can be calculated for all available covariates in a trial and ranked according to importance of inclusion in analysis to this end. This work seeks to address (1) the impact of failing to adjust for prognostic covariates at either randomization or analysis in a non-inferiority trial, (2) the development of the joint statistic for disparity and influence, and (3) the practical implications of its use as further outlined in the specific aims.

1.1 Motivating Example

This issue presented itself to the Neurologic Emergencies Treatment Trials (NETT) Network during the design of the Rapid Anticonvulsant Medication Prior to Arrival Trial (RAMPART). RAMPART is a non-inferiority trial in the pre-hospital treatment of status epilepticus, which is a life threatening and debilitating disease operationally defined as a seizure lasting for at least five minutes. The current FDA approved treatment for status epilepticus is intravenous (IV) lorazepam (4mg-10mg), but as it is difficult to establish an IV in a patient that is

actively seizing, there is a need for an equally effective therapy with a simpler method of administration. Intramuscular (IM) administration of a benzodiazepine had shown promise as an alternate approach, but was not expected to yield superior efficacy. As such, RAMPART was designed to test the hypothesis that IM midazolam (10 mg in adults or 5 mg in children < 13 kg) was non-inferior to IV lorazepam (4 mg in adults or 2 mg in children < 13 kg) by a pre-specified clinically meaningful amount, the non-inferiority margin. The primary outcome was the proportion of subjects with seizure termination prior to emergency department (ED) arrival without the use of rescue medication. A total of 893 adult and pediatric patients were enrolled.[40]

The RAMPART trial had many unique aspects including administration of treatment in a pre-hospital setting. This required extensive training of the Emergency Medical Services (EMS) units across the US and the need to have study drug available on each ambulance, making it infeasible to include prognostic factors in the randomization algorithm. Therefore, subjects were allocated via complete randomization, without consideration of any potentially important prognostic variables, and an unadjusted primary analysis was conducted.[40]

Upon completion of the trial, not only was IM midazolam found to be non-inferior to IV lorazepam ($\Pi_{IV} = 63.6\%,\ \Pi_{IM} = 73.4\%, p < 0.001$), but results suggested superiority (p<0.001), implying that a change in clinical practice

may be warranted.[39] There were, however, a number of covariates in this trial that may have been of prognostic importance, and the exclusion of these covariates from the analysis may have had serious implications in the accuracy of the estimation of the treatment effect as well as the overall conclusions regarding IM midazolam's non-inferiority to IV lorazepam.

1.2 Specific Aims

As the impact of covariate adjustment in the non-inferiority setting is less understood than in the context of superiority, it is of interest to evaluate the impact of failing to include covariates in the analysis of non-inferiority trials and to develop a method to aid in the decision to include covariates. Thus, the aims of this dissertation are:

1.) To evaluate differences in the operating characteristics (power, type I error, and bias in the treatment estimate as well as its accompanying standard error) in the superiority and non-inferiority settings under the following four scenarios:
    a. Complete randomization and unadjusted analysis
    b. Complete randomization and adjusted analysis
    c. Covariate adjusted permuted block randomization and unadjusted analysis

d. Covariate adjusted permuted block randomization and adjusted analysis

Hypotheses:

For the superiority setting, we hypothesize a reduction in type I error probability when covariate adjusted permuted block randomization is used with an unadjusted primary analysis (Scenario *1c*), and nominal type I error for all other scenarios. Furthermore, we expect to see greater preservation of power in adjusted analyses as compared to unadjusted analyses, with additional preservation occurring in the context of covariate adjusted randomization (Scenario 1*d*). As for bias, we hypothesize estimates of treatment effect and standard errors that are biased downward toward zero in the context of unadjusted analyses, with greater bias in the standard error of the treatment estimate when complete randomization is used. However, we expect minimal bias in the context of adjusted analyses.

In the non-inferiority setting, we expect a reversal such that type I error is inflated for Scenario *1c* (nominal for all other scenarios) and power is inflated for covariate adjusted randomization and analyses. Similarly, we hypothesize that treatment estimates will be biased upward toward zero and the accompanying standard error biased downward in the

context of unadjusted analysis, with greater bias in the standard error of the treatment estimate in the context of complete randomization. In addition, we expect minimal bias in the context of adjusted analyses.

2.) To develop a joint statistic to quantify the impact of adjustment for a given binary covariate based on its association with outcome (influence) and its imbalance across treatment groups (disparity)

Hypotheses:

We believe that the work of Canner[6,7] and Beach and Meier[3] can be extended for use in the non-inferiority framework with a binary outcome through the introduction of the non-inferiority margin. We expect that this statistic will fully quantify the impact of adjustment for a given covariate and can be used to rank available covariates in terms of relative importance. Finally, by choosing an appropriate subset of covariates to include in analysis based on this ranking, we believe that we can optimize reduction in type I error.

3.) To apply this statistic to rank the available covariates in a real trial application and evaluate the impact of adjustment on the estimate of treatment effect

Hypotheses:

We believe that, when applied to the RAMPART study as a whole and also to the pediatric subset, the joint statistics developed in Aim 2 will be able to identify the most appropriate subset of covariates to include in the analysis. We also believe that inclusion of these covariates will result in an estimate of treatment effect that is more precise and more conservative in terms of type I error.

Chapter 2

Review of Literature

2.1 Non-Inferiority Trials

Non-inferiority trials are typically used in situations in which a standard of care has already been established to be superior to placebo. However, a new treatment may provide potential advantages, such as a reduction in costs, greater ease of administration, or fewer side effects. In such a situation, this new candidate therapy may not be expected to be superior to standard of care, but if it alleviates some of the limitations and is "no worse" by some pre-specified amount, then it may be adopted by the clinical community. Thus, a non-inferiority trial seeks to provide a statistical answer to this question of "no worse than" by showing that the experimental treatment is non-inferior to the standard of care by some pre-specified amount, referred to as the non-inferiority margin (d). There are many challenges that arise in the design and analysis of such a trial including the proper specification of hypotheses, the determination of the probability of success in the standard of care group and the non-inferiority

margin, the appropriate analysis method to employ, and the impact of covariate adjustment.

A fundamental difference between a superiority trial and a non-inferiority trial is the way in which hypotheses are stated.[5] The typical null hypothesis for a one-sided superiority trial is that the difference is less than or equal to zero, whereas the alternative hypothesis is that the difference is greater than zero (experimental treatment is superior to placebo):

$$H_0 : \pi_T - \pi_P \leq 0 \ or \ \pi_T \leq \pi_P$$

$$H_1 : \pi_T - \pi_P > 0 \ or \ \pi_T > \pi_P$$

where π_T is the probability of success in the treatment group and π_P is the probability of success in the placebo group. In the framework of a non-inferiority trial, the null hypothesis is that the difference between the efficacy of the new treatment (π_T) and standard of care (π_S) is less than or equal to the non-inferiority margin (δ) versus the alternative hypothesis which states that this difference is greater than δ:

$$H_0 : \pi_T - \pi_S \leq -d \ or \ \pi_T \leq \pi_S - d$$

$$H_1 : \pi_T - \pi_S > -d \ or \ \pi_T > \pi_S - d$$

When the hypotheses are stated in terms of probabilities of success as above, the non-inferiority margin is often negative because we expect the new treatment to be less efficacious than the standard of care under the null.[5] However, if the

hypotheses are stated in terms of the proportion of failures in each treatment group, then we would expect a positive non-inferiority margin.

Regardless, we are not comparing the candidate treatment to a placebo as this would be unethical in the presence of a viable treatment. Instead, we are using the standard of care for our comparison.[5] This forces us to make an assumption about the probability of success in the standard of care group, which we can only do in light of previous studies in the same patient population. This is often referred to as the constancy assumption because we must assume that this treatment effect is constant across studies.[13,41] In addition, the lack of a placebo group forces us to make the assumption of assay sensitivity, which assumes that, had a placebo group been included, the standard of care would have shown this same treatment effect in comparison. If non-inferiority is demonstrated, this assumption can be evaluated, but never formally proven, through the comparison of confidence intervals for standard of care in the current study and in the historical study used to establish superiority over placebo.[13,41]

Another design challenge in a non-inferiority study is that of choosing the proper non-inferiority margin (δ). In 2006, the Committee for Proprietary Medicinal Products for Human Use (CHMP) guidelines[10] suggested that using an absolute difference in proportions of 10% is appropriate, whereas the FDA[41], borrowing from vaccine trials, proposed a stepwise function based on the probability of success for the standard of care (π_S) as follows:

$$\pi_S < 0.8 \rightarrow \delta = 0.20$$

$$0.8 \leq \pi_S < 0.9 \rightarrow \delta = 0.15$$

$$\pi_S \geq 0.9 \rightarrow \delta = 0.10$$

Romel[34] suggested a function that was similarly dependent on the reference treatment effect, but that was continuous:

$$-0.223\sqrt[3]{\{\pi_S\,(1-\pi_S\,)\}}$$

The current practice in the determination of the non-inferiority margin requires the investigator to place the experimental treatment on a continuum with the effect of placebo at one end and the effect of standard of care at the other. Then, the full distance from placebo to standard of care can be derived using the constancy assumption, and the distance from experimental treatment to standard of care, which is the non-inferiority margin, can be based on clinical relevance. Finally, the distance between placebo and experimental therapy can be calculated as[11]:

$$(1-\delta) * (\pi_S - \pi_P)$$

All of the methods discussed previously deal with a non-inferiority margin based on an absolute difference in proportions, a technique which Garrett[18] condemns for being dependent on the probability of success for standard of care, and offered the use of the odds ratio as a viable alternative that is not subject to this constraint. In this case, the margin corresponds to smaller absolute differences in proportions when the reference probability of success approaches either 0 or 1, which is

consistent with the goals of the regulatory documents, and it is easily incorporated into generalized linear models.

There has also been much debate in the statistical and clinical literature about the best approach to analyze data from a non-inferiority trial. In 1977, Dunnett and Gent[12] proposed the following test statistic based on the incorporation of the non-inferiority margin into the test for a difference in proportions in the context of a superiority hypothesis:

$$z = \frac{p_1 - p_2 - \delta}{\sqrt{var(p_1 - p_2)}}$$

where $var(p_1 - p_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$, p_i is the probability of success in treatment i, n_i is the number of patients in treatment i, and $i = 1,2$. Another hypothesis testing approach which has been proposed by Hung et al.[21] and endorsed by the FDA guidance[41] combines information from the trial(s) of the historical control versus placebo with information from the current non-inferiority trial as follows:

$$Z = \frac{log(\hat{T}) - log(\hat{C}) - (1-\lambda)(log(\widehat{C_O}) - log(\widehat{P_0}))}{\sqrt{\sigma_{TC}^2 + (1-\lambda)^2\sigma_{PC_O}^2}}$$

where $log(\hat{T}) - log(\hat{C})$ is the log relative risk from the non-inferiority trial of treatment versus standard of care with variance equal to σ_{TC}^2, $log(\hat{C}_O) - log(\widehat{P_0})$ is the log relative risk from the trial(s) of standard of care versus placebo

with variance equal to $\sigma^2_{PC_O}$, and λ is the percent of standard of care effect which must be preserved.

Several confidence interval approaches have also been proposed in which the lower bound of the two-sided 95% confidence interval for the effect of treatment versus standard of care is compared to either (1) the point estimate of the effect of standard of care versus placebo found in previous trials, (2) the upper bound of the two-sided 95% confidence interval of the effect of standard of care versus placebo, or (3) a pre-specified fixed margin of clinical significance. Either of the latter two methods results in conservative or nominal type I error, whereas the first can result in increased type I error due to the fact that the variance of the effect of standard of care versus placebo is not considered, and thus appropriate weight is not given to trials of different sizes.[21,41] In light of these findings, the current confidence interval approach is accomplished by calculating a two-sided 95% confidence interval of the risk difference, relative risk, or odds ratio of the treatment effect, and comparing its lower bound to a non-inferiority margin.[18] This technique is considered by many to be better than the typical hypothesis testing approach because it offers additional information such as the estimates of the minimum and maximum treatment effect.

The subject of covariate adjustment in non-inferiority trials remains lacking in the literature. Since the statement of null and alternative hypotheses yields a different implication for type I and type II errors when compared to the

framework of superiority, it is suspected that the impact of failure to adjust for important covariates may also differ in this setting. In non-inferiority trials, type I error (or α) is defined as claiming non-inferiority when the new treatment is actually worse than the active control, and type II error (or β) is defined as claiming lack of non-inferiority when the experimental treatment is non-inferior to the active control.[5] Because of this important difference, our understanding of covariate adjustment cannot easily be extended from the superiority to the non-inferiority setting. Thus, there is a great need to further evaluate the impact of covariate adjustment on operating characteristics in the context of non-inferiority trials.

2.2 Randomization

In clinical trials, the variability present in the outcome of interest is large in comparison to the expected magnitude of the treatment effect. The act of randomization seeks to control this variability through the elimination of selection bias as well as accidental bias in treatment allocation. Selection bias can be controlled by the introduction of randomness, which forms the basis for probabilistic approaches to hypothesis testing. However, randomization also seeks to control accidental bias by ensuring that there are no systematic differences between treatment groups.[14,19,30,44] The type of randomization that is

used affects the amount of resulting randomness and balance, which can impact the estimation of treatment effect if not considered at the analysis phase.[25,44]

In complete (or simple) randomization, patients are assigned to treatment groups with equal probability. This results in perfect randomness but can often lead to groups of unequal size, especially given small sample sizes. In fact, when we randomize *N* patients with probability *p=0.5* to two treatment groups, the probability that each group contains an equal number of patients is:

$$P[X = x] = \binom{N}{N/2}\frac{1}{2^N}$$

Thus, if N equals 100, the probability that randomization will yield exactly 50 patients per group is approximately 8%. Furthermore, given calculation of mean and variance from the binomial distribution, the same scenario yields a 95% confidence interval of the number of patients assigned to either Treatment A or Treatment B as roughly equal to:

$$N_{A\ or\ B} \pm 10$$

As such, we can expect more than a 60/40 imbalance in favor of one treatment over the other 5% of the time.[30]

Permuted block randomization, as introduced by Hill[19], is achieved by first choosing a block size, which can be any multiple of the number of treatments to be evaluated. Then, the order of treatment assignments within each block is randomly permuted, such that perfect balance is achieved after the completion of

each block. Thus, if a trial consisted of two treatments (*A* and *B*) as discussed previously, and a block size of two was chosen, treatment assignments within each block could take the form of either AB or BA. In this manner, the maximum imbalance that could result would be one, yet the proportion of deterministic assignments would be 50%.[30] Thus, permuted block randomization results in a much tighter treatment balance than complete randomization, but with much less randomness.[44] When the block size is equal to the sample size, permuted block randomization becomes a random allocation design. This results in near perfect randomness (some deterministic assignments may occur for the last subject(s) enrolled) as well as perfect balance at the conclusion of the trial.[44]

Any of the randomization strategies above can also be used in the presence of important covariates. A prognostic covariate is any variable that is significantly associated with outcome, and the exclusion of such factors from the randomization scheme or analysis may have important implications for a test of treatment effect. Thus, when there is a known prognostic covariate, adjustment is recommended.[1,22,29,31,36] This is often accomplished by first categorizing covariates, and then performing a given randomization scheme (e.g., complete randomization, random allocation, permuted block) within each stratum. In so doing, balance is achieved not only with respect to the number of patients within each treatment group, but also with respect to the distribution of covariates within these groups. This covariate adjusted randomization imposes additional balance,

which yields greater power in detecting treatment effects, serves to protect against model misspecification, and ensures greater power for subgroup and interim analyses.[22,25,35,38]

2.3 The Impact of Failing to Adjust for Covariates in Analysis

It has been shown in a superiority setting that failure to adjust for important prognostic covariates at analysis leads to biased estimates of the treatment effect, the direction and magnitude of which is dependent on the strength of association between the covariate and outcome, as well as the level of covariate imbalance across treatment arms.[9,16,17,23,26,29,33,35,38]

Robinson and Jewell[33] show that in a linear setting, a strong association between covariate and outcome decreases the variance of the treatment effect in the context of adjustment through the reduction of residual variance, whereas a strong association between covariate and treatment increases the variance of the treatment effect. Thus, precision is based on these competing effects. As for logistic regression, both a strong association between covariate and outcome and a strong association between covariate and treatment result in increased variance given adjustment.[33] Therefore, in logistic regression, there is always an automatic loss in precision when we adjust for covariates, regardless of whether or not they are predictive of outcome.[26] Thus, the standard error of the treatment effect in an unadjusted analysis will be deflated as compared to the standard error in the

adjusted analysis. However, it is important to note that this bias in the standard error of the estimate of the treatment effect is coupled with a bias in the treatment estimate itself.[33]

Gail et al.[17] illustrate the nature of this bias in the treatment estimate via Taylor series approximation in a nonlinear setting. In a linear setting, the treatment effects in an adjusted versus an unadjusted model are equivalent. In the nonlinear setting (i.e., in the context of logistic regression), they quantify the discrepancy between the unadjusted treatment effect estimate and the adjusted treatment effect estimate as:

$$\beta_{trt}{}^{*} - \beta_{trt} = \frac{1}{2}\beta_{cov}^{2}\sigma_{cov}^{2}\left[\frac{h''(\beta_{trt,})}{h'(\beta_{trt})} - \frac{h''(\beta_0)}{h'(\beta_0)}\right]$$

where $\beta_{trt}{}^{*}$ is the treatment effect when the covariate is not included, β_{trt} is the treatment effect when the covariate is included, β_{cov} is the covariate effect, σ_{cov}^{2} is the variance of the covariate, $h(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$, $\eta = \beta_0 + \beta_{trt}(Trt) + \beta_{cov}(Cov)$, and β_0 is the intercept. They demonstrate that this discrepancy is nonzero (i.e., the unadjusted effect will always be biased) except when there is no treatment effect, when there is no association between covariate and outcome, or when the variance of the covariate is zero. Furthermore, Gail et al.[17] show that the bias tends to be such that the unadjusted effect will underestimate the adjusted effect if the treatment effect is positive, and this outweighs any benefit in standard error. Thus, the overall result is a decrease in power when failing to adjust, and this

effect holds even when important prognostic covariates are perfectly balanced.[3,15-17,26,33]

In addition to the strength of association between covariate and outcome and covariate and treatment, the marginal distribution of the treatment itself also plays a role in the variance of the treatment estimate in logistic regression such that the highest variance is achieved when the probability of success is 0.5. For this reason, the operating characteristics derived from analyses of a clinical trial may be sensitive to assumptions regarding the probability of success in the control arm and the minimum clinical difference.

The use of covariate adjusted randomization in the design of a clinical trial introduces correlation between treatment and covariate. As presented by Kahan and Morris[23], in the context of superiority, the variance of the treatment difference within a stratum, given a continuous outcome, can be found from the following expression:

$$var(\bar{Y}_1 - \bar{Y}_2) = var(\bar{Y}_1) + var(\bar{Y}_2) - 2cov(\bar{Y}_1, \bar{Y}_2)$$

In randomization schemes that do not use covariate information, the mean outcomes for treatment one ($\bar{Y}_1$) and treatment two ($\bar{Y}_2$) can be assumed to be independent. Therefore, the last term would be equal to zero, and the previous equation, given equal variance of the treatment groups, would simplify to:

$$var(\bar{Y}_1 - \bar{Y}_2) = \frac{2\sigma^2}{n}$$

However, in the absence of this assumption of independence, the correlation between $\bar{Y}_1$ and $\bar{Y}_2$ is ρ, and the following is applicable:

$$var(\bar{Y}_1 - \bar{Y}_2) = \frac{2\sigma^2}{n} - \frac{2\rho\sigma^2}{n} = \frac{2\sigma^2}{n}(1-\rho)$$

Thus, if we were to conduct an unadjusted analysis in the context of covariate adjusted randomization, the estimated variance of the treatment effect would be biased upwards, leading to decreased power and type I error when the outcome is continuous.[23,29,38] The magnitude of this bias is directly associated with not only the strength of association between covariate and outcome, but also with the strength of association between covariate and treatment (i.e., the level of confounding). [1,31,35]

In light of all of the statistical evidence in favor of adjusted analyses, it is currently recommended that important prognostic covariates that are included in the randomization scheme should also be included in the final analysis in the form of a properly specified analysis of covariance.[1,22,29,31,36] Whereas failure to adjust for covariates in the superiority setting decreases type I error and power as described previously, failure to adjust in non-inferiority may actually increase type I error and power, but there has only been one paper to date on this subject.[18] Thus, the impact of covariate adjustment in the context of non-inferiority is an important issue worthy of further investigation.

2.4 The Impact of Tests of Baseline Imbalance

Given that the impact of failing to adjust for important prognostic covariates is dependent on both the strength of association with outcome as well as the level of imbalance, it has become common practice among some clinical trialists to perform statistical tests of baseline covariates, and then to adjust for those covariates which are significantly imbalanced. Although this practice has been condemned in the statistical literature[1,31,32,35,36], it remains prevalent, the rationale being that treatment groups should be as similar as possible for an unbiased assessment of the treatment effect. According to Altman and Dore[2], 46 of the 80 trials published in four leading medical journals in 1987 conducted tests of baseline imbalance resulting in roughly 600 tests, of which only 24 (4%) were significant at the 5% level Pocock et al.[31] find similar trends in 2002. Twenty-four (24) out of the 50 trials that they investigated (48%) employed such tests, of which 18 out of 299 total tests (6%) were significant. Thus, as Altman[1] eloquently states, "performing significance tests to assess baseline variables is to assess the probably of something having occurred by chance when we know that it did occur by chance." As a result, this practice serves as merely a test of the process of randomization itself. Balance is, in fact, not necessary for valid inference, but rather concerns the precision of the treatment effect. Furthermore, the exclusion of a highly influential covariate even if it has only a very small (statistically insignificant) imbalance across treatment arms will influence the final analysis,

whereas the exclusion of a highly imbalanced covariate with no association with outcome will have no effect.[1,31,36]

The practice of testing for baseline imbalance also has important implications for the type I error and power of the final analysis. Following notation given in an explanation by Senn[35] using a continuous outcome and a single continuous covariate, an unadjusted test of treatment effect under the null hypothesis in superiority is as follows:

$$\frac{d_y}{\sqrt{\sigma_y^2\left(\frac{1}{n_{trt}}+\frac{1}{n_{cont}}\right)}} \geq Z_\alpha$$

where d_y is the treatment effect, σ_y^2 is the variance of the treatment effect, n_{trt} and n_{cont} are the sample sizes of the treatment and control groups respectively, and Z_α is the (1-α)x100% standard normal quantile. Senn derives the following expression for the size of a test of treatment effect conditional on a covariate X:

$$\alpha(d_x) = p\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho d_x}{\sqrt{(1-\rho^2\sigma_x^2\left(\frac{1}{n_{trt}}+\frac{1}{n_{cont}}\right)}}\right]$$

where $\alpha(d_x)$ is the conditional size of the test, d_x is the effect of the covariate, σ_x^2 is the variance of the covariate effect, and ρ is the strength of the association between covariate and outcome. Finally, if we let d_x^* represent Senn's "standardized prognostic imbalance" between treatment groups (i.e., the two

sample Z statistic comparing mean covariate values across treatment groups), this reduces to:

$$\alpha(d_x) = p\left[Z \geq \frac{Z_\alpha}{\sqrt{1-\rho^2}} - \frac{\rho d_x^*}{\sqrt{(1-\rho^2)}}\right]$$

From this, we can see that for a given strength of the covariate (ρ), as the level of imbalance increases the size of the test also increases in a non-linear fashion in the superiority setting. Furthermore, when the covariate is negatively associated with outcome, increasing levels of imbalance result in increasingly conservative unadjusted tests. Similarly, when both the covariate and the treatment are positively associated with outcome, an imbalance favoring the treatment group will increase power, the magnitude of which will increase as the strength of the covariate increases.[35] On the other hand, when a covariate is negatively associated with outcome and the treatment effect is positive, an imbalance favoring the treatment group will result in decreased power.[8] Finally, it is important to note that such effects on type I error and power occur prior to the cutoff for significance (5%) of a baseline test of imbalance. Thus, not only does this practice impact the error probabilities of the final analysis, but is also offers no guarantee that the covariate imbalance will not influence overall conclusions.[1,3,8,29,35,36] However, this practice of testing for baseline imbalance, much like covariate adjustment in general, has received little attention in the non-inferiority setting, and thus it is a topic worthy of further investigation.

2.5 Quantifying the Impact of Adjustment for a Given Covariate Based on Influence and Disparity

Much of the regulatory and statistical literature agrees that covariate adjustment should be specified *a priori* and that those covariates included in randomization should also be included in the final analysis.[22,36] However, methods for identifying important covariates that are not known in advance are lacking. While significance testing continues to be a prominent practice, an alternative has been proposed that may not have the same impact on type I error.

As previously stated, the impact of including a given covariate in analysis can be quantified in terms of its association with treatment and its association with outcome. In practice, these correlations are often evaluated in isolation despite the fact that they represent competing goals. However, the simultaneous evaluation of these factors can reveal the optimal estimate of the treatment effect in terms of both the point estimate itself and its precision.

At the Society for Clinical Trials Annual Scientific Sessions in 1981, Canner[6] presented such a method for the inclusion of a binary covariate given a binary outcome and zero treatment effect, which was later included in a paper by Beach and Meier[3]. They present the following condensed 2x2x2 table in which the number in each cell can be obtained by multiplying the proportions of failure (top expression) by the total number of subjects (bottom expression).

Table 2.5.1 Condensed 2x2x2 Table of Outcome, Treatment, and Covariate

	Cov=0	Cov=1	
Control	$p(1-r_2)$ $N(1+r_1)/2$	$p(1+r_2)$ $N(1-r_1)/2$	$p(1-r_1r_2)$ N
Trt	$p(1-r_2)$ $N(1-r_1)/2$	$p(1+r_2)$ $N(1+r_1)/2$	$p(1+r_1r_2)$ N
	$p(1-r_2)$ N	$p(1+r_2)$ N	p $2N$

where N is the number per group, p is the proportion of failures, $r_1$is the correlation between treatment and covariate,and r_2 can be derived as follows:

$$r_2 = \frac{corr(cov, out)\sqrt{p(1-p)}}{p}.$$

From this, a statistic for the association between covariate and outcome (Z_I), which Canner calls "influence", was developed in which the numerator represents the marginal difference between the overall proportion of failures among those with the covariate verses those without and the denominator represents its accompanying variance.

$$Z_I = \frac{p(1+r_2) - p(1-r_2)}{\sqrt{2pq/N}} = \frac{\sqrt{2N}pr_2}{\sqrt{pq}}$$

where $q = 1 - p$. He similarly derives a statistic for the association between covariate and treatment (Z_D), which he refers to as "disparity", based on the number in the treatment group with the covariate minus the number in the treatment group without the covariate.

$$Z_D = \frac{\left[\left(\frac{N}{2}\right)(1+r_1) - (\frac{N}{2})(1-r_1)\right]\sqrt{2N}}{\sqrt{N^2}}$$

Finally, he combines these to produce a joint statistic (Z_U) which gives equal weight to disparity and influence.[3,6]

$$Z_U = \frac{Z_I\, Z_D}{\sqrt{2N}} = \frac{p(1+r_1r_2) - p(1-r_1r_2)}{\sqrt{2pq/N}} = \frac{\sqrt{2N}pr_1r_2}{\sqrt{pq}}$$

This Z_U represents the importance of adjustment for a given covariate, and when calculated for all candidate covariates in a dataset and ranked, it can be used to evaluate their relative importance. While this joint statistic is of great theoretical importance, it suffers from the impractical assumption that there is no association between treatment and outcome (i.e., the adjusted statistic, Z_A, is equal to zero). In order to accommodate non-zero treatment effect, a third layer of correlation must be added. Beach and Meier[3] provide this extension in the context of continuous outcome in terms of regression parameters, which Canner[7] later presents via correlation.

$$Z_A = \frac{b_{y1.2}}{\sqrt{\widehat{var b_{y1.2}}}} = \frac{r_{y1.2}\sqrt{N-3}}{\sqrt{1-r_{y1.2}^2}}$$

where $b_{y1.2}$ is the partial regression coefficient for outcome (Y) and treatment (X_1) given covariate (X_2) and $r_{y1.2}$ is the partial correlation coefficient for outcome and treatment given covariate. This can be expanded as follows:

$$Z_A = \frac{(r_{y1} - r_{y2}r_{12}\sqrt{N-3})}{\sqrt{1 - r_{y1}^2 - r_{y2}^2 - r_{12}^2 + 2r_{y1}r_{y2}r_{12}}}$$

given the well- known relationship

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}$$

where $r_{y1}$is the correlation between outcome and treatment, r_{y2} is the correlation between outcome and covariate, and r_{12} is the correlation between treatment and outcome. If we subtract Z_A from Z_U, we can obtain a value for the importance of adjustment controlling for the treatment effect.[7]

These methods are sufficient to fully quantify the impact of adjustment for a single covariate or multiple covariates given that they are independent of each other. However, in a real trial setting, a strictly additive model is rarely sufficient. Although theoretical evaluations of this scenario become increasingly impractical as the number of important covariates increases, Berger[4] proposes an alternate method for the evaluation of disparity in which the first covariate is selected in the usual way, but subsequent covariates are conditioned such that the adjusted discrepancy is the weighted average of the disparity in each of the strata formed by the existing covariates.

The impact on type I error of these data driven methods for covariate selection is not well understood even in the context of superiority as they have not been well adopted by the clinical trials community. In non-inferiority, it has been

shown that failure to include important covariates increases type I error.[18,27] Thus, these methods are worth further evaluation. The following three chapters consist of three original papers designed to build upon previous research in the areas of covariate adjustment at randomization and analysis, methods for covariate selection when prognostic factors are not known *a priori*, and practical application in accordance with the three aims as previously stated.

Chapter 3

Aim 1:

The Impact of Covariate Adjustment at Randomization and Analysis for Binary Outcomes: Understanding Differences Between Superiority and Non-Inferiority Trials

The question of when to adjust for important prognostic covariates often arises in the design of clinical trials, and there remain various opinions on whether to adjust during both randomization and analysis, at randomization alone, or at analysis alone. Furthermore, little is known about the impact of covariate adjustment in the context of non-inferiority designs. The current simulation-based research explores this issue in the non-inferiority setting, as compared to the typical superiority setting, by assessing the differential impact on power, type I error, and bias in the treatment estimate as well as its standard error, in the context of logistic regression under both complete and covariate adjusted permuted block randomization algorithms.

In both the superiority and non-inferiority settings, failure to adjust for covariates that influence outcome in the analysis phase, regardless of prior

adjustment at randomization, results in treatment estimates that are biased toward zero, with standard errors that are deflated. However, as no treatment difference is approached under the null hypothesis in superiority and under the alternative in non-inferiority, this results in decreased power and nominal or conservative (deflated) type I error in the context of superiority, but inflated power and type I error under non-inferiority. Results from the simulation study suggest that, regardless of the use of the covariate in randomization, it is appropriate to adjust for important prognostic covariates in analysis, as this yields nearly unbiased estimates of treatment as well as nominal type I error.

3.1 Introduction

The non-inferiority trial design is growing progressively more popular as the need for comparable therapies with secondary advantages increases. The challenges that arise in the design and analysis of such a trial have been discussed in the FDA Guidance on non-inferiority trials[41] as well as methodological research on proper specification of hypotheses, the choice of active control, determination of the non-inferiority margin, and the appropriate analysis method.[5,11-13,18,22] However, critical gaps in the literature remain regarding key design issues, specifically the impact of covariate adjustment at both randomization and at the analysis phase.

In a PubMed search of phase-III clinical trials published in the last ten years containing the word "non-inferiority", 44.8% of 58 trials utilized complete randomization and unadjusted primary analyses. Covariate adjusted randomization and appropriate analysis of covariance or stratification was used in 20.7% of trials. Covariate adjustment at randomization, but not at analysis was done in 25.7%, and adjustment at analysis but not at randomization was done in 8.6% of the trials. This review of current practice shows that there is still considerable variability in the way that clinical trialists use covariate information, a finding that is not surprising considering the lack of attention it has received in the non-inferiority literature.

This chapter expands the research on covariate adjustment in the non-inferiority setting by examining the impact of adjustment at randomization as well as at analysis using logistic regression models. A simulation study is conducted to examine the operating characteristics in both superiority and non-inferiority settings. Section 1.2 reviews the existing statistical literature on covariate adjustment. Sections 1.3 and 1.4 present the simulation methods and results, and Section 1.5 discusses the differential impact of covariate adjustment in these two settings.

3.2 Existing Literature

Little work has been published to date to examine the impact of covariate adjustment at both randomization and analysis in the context of non-inferiority trials. Garrett[18] explains that, due to the reversal of hypotheses from the superiority setting, there is also a reversal of the impact of errors. Thus, he cautions readers that the power and type I error is inflated when important prognostic factors are ignored in the non-inferiority setting. However, this work does not take into account the potential impact of the randomization scheme.

It has been shown in a superiority setting that failure to adjust for important prognostic covariates at either randomization or analysis leads to biased estimates of the treatment effect, the direction and magnitude of which is dependent on the strength of association between the covariate and outcome, as well as the level of covariate imbalance across treatment arms.[9,16,17,23,26,29,33,35,38] Gail et al.[17] illustrate the nature of this bias via Taylor series approximation in a nonlinear setting. They show that in a linear setting, the estimated treatment effect in an adjusted versus in an unadjusted model are equivalent. In the nonlinear (i.e., the logistic) setting, they quantify the discrepancy between the adjusted treatment effect estimate and the unadjusted treatment effect estimate and demonstrate that this discrepancy is nonzero (i.e., the unadjusted effect will always be biased) except when there is no treatment effect, when there is no association between covariate and outcome, or when the variance of the covariate

is zero. Furthermore, Gail et al. show that the bias tends to be such that the unadjusted effect will underestimate the adjusted effect if the treatment effect is positive, resulting in a decrease in power, and that this effect holds even when covariates are perfectly balanced. As Robinson and Jewell[33] point out, this underestimation of the treatment effect when failing to adjust for covariates outweighs any benefit in standard error.[3,15,16,17,26,33]

The impact of covariate adjustment at randomization has been evaluated in the context of superiority. Kahan and Morris[23] illustrate that, for continuous, binary, and time-to-event outcomes, stratified randomization creates correlation between treatment arms. Thus, unadjusted analyses in this context result in decreased power and type I error, as well as inflated standard errors of the treatment effect.

In light of the statistical evidence in favor of adjusted analyses, it is clear that important prognostic covariates that are included in the randomization scheme should also be included in the final analysis in the form of a properly specified analysis of covariance,[22,29,36] and, in fact, this results in unbiased estimates of treatment effect, as well as nominal power and type I error. However, the full impact of covariate adjustment in the non-inferiority setting remains to be demonstrated and is a topic worthy of further investigation.

3.3 Simulation Studies

Parameters are specified according to both the hypothesis of interest (null or alternative) and the scenario of interest (superiority or non-inferiority). Four simulation studies were designed to perform complete randomization and covariate adjusted permuted block randomization in the context of superiority and non-inferiority designs. All simulations conducted both unadjusted and adjusted analyses based on the following models:

$$Unadjusted: \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 Trt$$

$$Adjusted: \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 Trt + \beta_2 Covariate$$

The probability of success in the control group (π_C) was set at 80%. This value was chosen based on a non-inferiority setting where the active control may have a high probability of success however the experimental treatment offers other advantages such as fewer side effects.

3.3a Simulation Parameters

The probability of success in the treatment group (π_T), the sample size, and the pre-determined margin were differentially specified depending on the statistical hypothesis to be tested (superiority or non-inferiority), but the simulation strategy remained the same. For the superiority setting, the probability of success in the treatment group was set at 90% in order to mimic a trial with an

expected absolute difference in treatment of 10%. The null hypothesis is $H_0: \pi_C = \pi_T = 0.80$, and the alternative is $\pi_C = 0.80$ and $\pi_T = 0.90$. Thus, β_1 is estimated to be $\ln(1) = 0$ under the null hypothesis and $\ln(2.25) = 0.811$ under the alternative. This information was then used to calculate the total sample size of $N = 392$ subjects in order to ensure power of 80% when there was no effect of the covariate ($\beta_2 = 0$).

The value of $\beta_2$was allowed to vary from -3.0 to 3.0 in increments of 0.5 in order to evaluate the effect of strength of covariate on operating characteristics, and equations (1) and (2) above were used to derive the intercept, β_0, for each scenario. Although it is not expected that any covariate would have a coefficient as high as 3.0, this range was included in the simulation to examine the theoretical setting.

For the case of non-inferiority, we set the non-inferiority margin at 0.10, resulting in the following hypotheses:

$$H_0: \pi_C - \pi_T \geq 0.10$$

$$H_1: \pi_C - \pi_T < 0.10$$

where $\pi_C = 0.80$ and $\pi_T = 0.70$ under the null and $\pi_C = \pi_T = 0.80$ under the alternative. This yielded a $\beta_1$under the null of -0.539 and under the alternative a value of zero. An odds ratio of 0.583 was derived from the expected probability of success in each of the two groups, which served as the cutoff for claiming non-inferiority. The values of β_2 and β_0 remained similar to the superiority setting.

The total sample size was estimated as N=676, so that the power to detect a treatment difference when $\beta_2 = 0$ was again 80%.

It should be noted that results on the linear/risk difference scale are not necessarily immediately applicable to the logistic/odds ratio scale but may require translation. The authors present hypotheses and parameters in terms of the linear scale to facilitate communication with clinical investigators but then analyze results using logistic regression to avoid convergence issues that would otherwise arise in the tails of the nonlinear distribution.[26] It is also noteworthy that sample size has to be increased slightly when the effect size is translated from a risk difference to an odds ratio.[20]

3.3b Simulation Strategy

The simulated subject dataset is filled sequentially by first assigning the level of covariate (0 or 1), based on dichotomization of a random uniform distribution, and then a treatment indicator via either complete randomization, where the probability that the i[th] patient is assigned to treatment ($p_{i,trt}$) is 0.5, or permuted block randomization within each level of the covariate according to the following:

$$p_{i,trt} = \frac{\left(b/2\right)\left(1 + int\left((i-1)/b\right)\right) - n_{i-1,trt}}{b\left(1 - int\left((i-1)/b\right)\right) - (i-1)}$$

where i = subject, b = block size, *int* = next highest integer value, and n_{i-1} = number previously assigned to treatment. A block size of six was chosen as a compromise between the authors' beliefs about current popular practice and the desire for results to be comparable to those of Kahan and Morris[23], who used a block size of eight. The probability of success for each patient is assigned as follows:

$$p_{i,\ success} = \frac{\exp(\beta_1 Trt + \beta_2 Covariate)}{1 + \exp(\beta_1 Trt + \beta_2 Covariate)}$$

and again compared to a random uniform distribution for dichotomization. Once the subject table is populated, unadjusted and adjusted analyses are conducted, and estimates of the odds ratios of the treatment effect, as well as their standard errors and two-sided 95% confidence intervals are extracted.

Power, defined as the percentage of trials under the alternative hypothesis in which the lower bound of the confidence interval for the odds ratio crosses the pre-determined margin (>1.0 for superiority or >0.583 for non-inferiority), as well as type I error, defined as the percentage of trials under the null in which the lower bound of the confidence interval for the odds ratio crosses the pre-determined margin, are calculated across 10,000 iterations. Bias in the estimate of the treatment effect, defined as the average difference between the estimate and the true value of β_1, and bias in the standard error of the estimate of treatment effect, defined as the mean difference between empirical and model standard

errors, are calculated under the alternative hypothesis in the context of superiority and under the null hypothesis in the context of non-inferiority, as this is where one would expect to find a treatment difference.

3.4 Simulation Results

3.4a Superiority

In the context of superiority, there is a slight loss in power as β_2 (which represents the strength of the covariate) moves away from zero in either direction for unadjusted analyses and a slight gain in power for adjusted analyses. Furthermore, there appears to be little impact of balancing at randomization among adjusted analyses, but some improvement of covariate adjusted permuted block randomization over complete randomization among unadjusted analyses. Type I error, under all scenarios, yields close to nominal values regardless of the value for β_2, with the exception of the scenario that employed covariate adjusted permuted block randomization coupled with an unadjusted analysis. In this scenario, type I error decreased as β_2 moved away from zero in either direction.

Under the alternative hypothesis, we can see that the unadjusted analyses' treatment effects and standard errors underestimate those of the adjusted analyses. Furthermore, the bias in the standard error appears to be less severe for covariate adjusted permuted block randomization than for complete randomization due to the correlation it creates between treatment groups. However, the treatment

estimate, as well as its accompanying standard error, is nearly unbiased (a slight positive bias was detected, but determined to be minimal) in adjusted analyses, with negligible effect of covariate balancing at randomization.

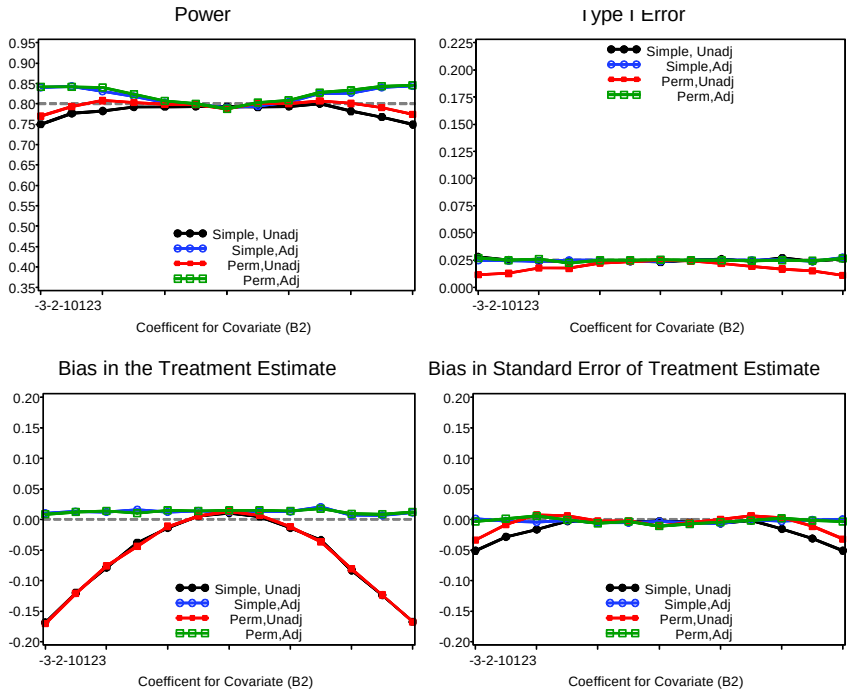

Figure 3.4a.1 Operating Characteristics for Covariate Adjustment in Superiority Trials

3.4b Non-Inferiority

The operating characteristics in the context of non-inferiority are quite different from superiority. In this setting, power decreases slightly as β_2 moves away from zero in either direction for adjusted analyses, and type I error is nearly

maintained regardless of covariate balancing at randomization via permuted block. For unadjusted analyses, an opposite effect from that demonstrated in the context of unadjusted analyses in superiority is observed. Power and type I error is increased as β_2 moves away from zero in either direction. This increase in type I error for unadjusted analyses in the presence of an influential prognostic covariate is greater in complete randomization than in the context of covariate adjusted permuted block randomization. Type I error rates are nearly maintained for adjusted analyses regardless of balancing at randomization.

The treatment estimate and its standard error are unbiased for adjusted analyses, whereas the treatment estimate is biased upward toward zero for unadjusted analyses regardless of balancing at randomization via permuted block. The standard error of this estimate follows the same pattern as the alternative hypothesis in the superiority setting (namely, standard errors are deflated for unadjusted analyses with covariate adjusted permuted block randomization yielding a less pronounced effect).

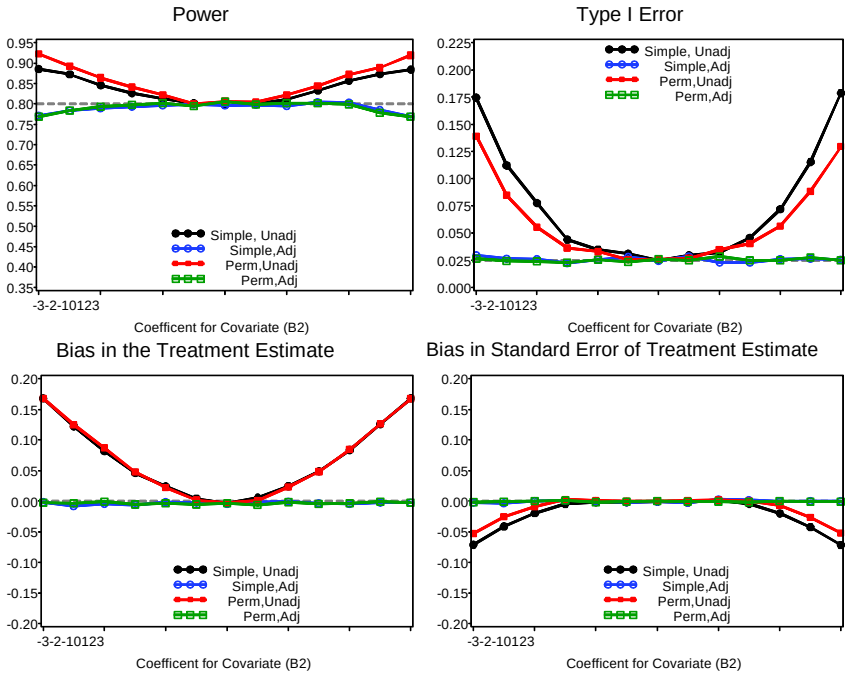

Figure 3.4b.1 Operating Characteristics for Covariate Adjustment in Non-Inferiority Trials

3.5 Discussion

The results presented above demonstrate that adjustment for important prognostic covariates in analysis is always preferred, regardless of whether the hypothesis to be tested is one of superiority or non-inferiority. In a superiority setting, adjusted analyses yield greater power to detect a treatment difference as compared to unadjusted analyses, as well as nominal type I error. In a non-inferiority setting, adjusted analyses do not have the added benefit of increased

power but are protective against unacceptable increases in type I error. Furthermore, adjusted analyses yield nearly unbiased estimates of treatment effect in both scenarios and biases in standard error that seem unavoidable given the nature of the nonlinear relationship between outcome and predictors.

These findings expand upon those of Garrett[18] by further quantifying the impact of failure to adjust for important prognostic covariates in terms of bias in the treatment estimate and subsequent implications for power and type I error. It also provides noteworthy extensions in terms of the evaluation of different covariate adjustment strategies (i.e., randomization versus analysis), a direct comparison with a superiority design, and the extendibility of such findings to different specifications of parameters. For example, in addition to the strength of association between covariate and outcome, the probability of success in the control also plays a role in the variance of the treatment estimate such that different starting points do not translate to equivalent odds ratios. Thus, an absolute non-inferiority margin of 0.10, which we set in the non-inferiority setting, does not always translate to a relative non-inferiority margin of 0.583, but only given an assumed probability of success in the control of 80%. To demonstrate the sensitivity of findings, a second simulation was undertaken in which the absolute non-inferiority margin remained 0.10, but the probabilities of success were changed to $\pi_C = 0.55$ and $\pi_T = 0.45$ (which translates to a relative non-inferiority margin of 0.669). In this scenario, the overall variability is

increased and the impact of failure to adjust for covariates is magnified, but similar trends were observed in all operating characteristics for the non-inferiority setting.

A limitation of this work lies in the fact that the observed differences between unadjusted and adjusted analyses can be thought of in terms of model inconsistency. The adjusted model is a conditional, or subject specific model, whereas the unadjusted model is a marginal, or population averaged, model. Thus, one could argue that the bias illustrated in the unadjusted treatment effect estimates is actually just a result of the difference between the two modeling strategies.[24,37] While this is worth comment, the simulation results, as presented, retain their practical value as an explanation of the impact of covariate adjustment.

Finally, the current research suggests that, in the context of adjusted analyses of binary outcomes, covariate adjusted permuted block randomization may not provide much gain over complete randomization. In fact, there were only minimal effects of balancing at randomization in the context of adjusted analyses for any of the operating characteristics that were presented, a result that was maintained even under a scenario in which a significant 70/30 covariate imbalance was forced (simulation not shown). In addition, the use of covariate adjusted permuted block randomization results in a loss of randomness via an increase in the number of deterministic assignments.[44] However, further research

is required to determine whether these findings can be generalized to other randomization schemes.

This research illustrates the impact of covariate adjustment at randomization and analysis in non-inferiority trials with binary outcomes and demonstrates the importance of conducting adjusted analyses in the presence of important prognostic covariates. Although we have exaggerated the covariate effect (i.e., β_2), this was done for purely theoretical purposes. In practice, it is unlikely that one would see a covariate effect of +/- 3.0 and the resulting type I error rate of 0.175. Regardless, our results hold as we see an inflation in the type I error when the covariate is not equal to 0; for example, a more practical case of a covariate effect of +/- 1.5 shows a type I error rate of 0.045. This suggests that proper adjustment will result in unbiased estimates of treatment and reduce the probability of committing a type I error, which is of particular importance in non-inferiority trials, as claiming non-inferiority when it is false may have severe implications for patients.

Chapter 4

Aim 2:

Choosing Covariates for Adjustment in Non-Inferiority Trials with Binary Outcomes Based on Influence and Disparity

It has been shown that the type I error rate is inflated when important covariates are excluded from a non-inferiority analysis.[18,27] Traditionally, whether to adjust for a covariate in a model is based solely on statistical significance or some other criteria such as Akaike information criterion (AIC) that relates to the magnitude of the effect on outcome. In addition, some select based on tests of baseline imbalance. However, several authors suggest that these aspects be considered simultaneously. Canner[6] developed a statistic for binary outcomes to determine the relative importance of including a covariate in a model based on both its effect on outcome and its association with treatment. Although Canner's approach assumed no treatment effect, Beach and Meier[3] extended this to non-zero treatment effects in the context of linear regression. The current research combines the methods of Canner for binary outcomes with the methods of Beach and Meier for non-zero treatment effect in order to quantify the relative

importance of including covariates in a non-inferiority study with a binary outcome. Theoretical results are presented and applied via simulation.

4.1 Introduction

In observational studies, the goal of covariate adjustment is often to quantify the impact of various baseline factors on outcome in order to understand the predictors of disease. In large randomized clinical trials, the goal becomes the unbiased and precise estimation of a treatment's effect on outcome. A truly unbiased estimate of treatment effect results from adjustment for any covariate with even a slight impact on outcome. As this is, of course, impossible in practice, a method of choosing the most important subset of potential covariates is of value.

If the goal is to obtain an accurate and unbiased estimate of the treatment effect, then one should choose a subset of covariates based on their association with outcome. This can be done via tests of bivariate association in which any covariate reaching statistical significance is included in the model for outcome. Alternatively, stepwise selection procedures which consider the iterative addition or removal of a single, or subset, of covariates can be employed, and decisions can be based on p-values, AIC, or some other criteria of influence. For binary outcomes, accuracy and precision represent competing goals; as the number of covariates in the model increases, the accuracy of the point estimate will also

increase, but the precision will decrease given the non-linear nature of the relationship with outcome.

If precision of the treatment estimate is the goal, then one should choose covariates based on the disparity with treatment allocation rather than association with outcome. Some trialists do not consider this issue in the formation of the model, but rather trust that the randomization scheme has controlled for any chance imbalances.[1] Others choose to test this formally with two-sample tests for baseline imbalance and then adjust for those covariates reaching statistical significance. While this practice has been condemned in much of the regulatory and statistical literature due to its impact on type I error, [22,36] it remains popular and is often reported in the first table of many manuscripts.

Perhaps, a better approach for choosing covariates would consider both influence on outcome and disparity with treatment. In a presentation to the Society for Clinical Trials in 1981, Canner[6] developed a statistic to do this simultaneously for binary outcomes with no treatment effect, which Beach and Meier[3] later extended for continuous outcomes with non-zero treatment effect. The current research seeks to expand Canner and Beach and Meier's work for application to a non-inferiority design with a binary outcome.

A typical superiority design generally seeks to show that a new drug is superior to placebo by some predetermined amount. A non-inferiority study, however, is often used when previous studies have established superiority of a

therapy, but a new therapy may offer some additional benefit such as lower cost, greater ease of administration, or fewer side effects. In this case, if the new drug is equally, or even slightly less, efficacious than the current standard of care, it would likely be of clinical benefit. Thus, the non-inferiority design seeks to answer the question of "no worse than" by some amount, the non-inferiority margin (d), that would be deemed an acceptable loss given the benefit tradeoff. This reversal of hypotheses leads to a reversal in the implications of type I and type II error.[5] As such, biased and imprecise estimates of treatment effect in non-inferiority studies can increase the likelihood of rejecting the null hypothesis and claiming non-inferiority. Thus, there is often a non-trivial inflation of the type I error probability when analyses fail to account for important covariates.[18,27] Given this important difference, there is potential for an extension of Canner and Beach and Meier's joint statistic for influence and disparity for use in identification of important covariates requiring adjustment in non-inferiority trials. Section 4.2 reviews the existing literature. Section 4.3 presents a theoretical argument for the inclusion of the non-inferiority margin in the statistic, followed by a simulation study in Sections 4.4 through 4.6, advice for practical use in Section 4.7, and discussion in Section 4.8 .

4.2 Existing Literature

Canner[6] presented a method to quantify the impact of adjustment for a binary covariate given a binary outcome and no treatment effect that was based on both disparity (i.e., imbalance between treatment arms) and influence on outcome. In 1989, Beach and Meier[3] expanded upon Canner's result to include a non-zero treatment effect. They presented the following condensed 2x2x2 table, in which the columns corresponding to the number of successes for each combination of treatment and covariate have been removed.

Let N be the number per treatment arm, p be the overall proportion of failure, and $r_1$be the correlation between treatment (X_1) and covariate (X_2). r_2 quantifies the strength of the association between covariate (X_2) and outcome (Y) and can be derived as follows:

$$r_2 = \frac{corr(X_2, Y)\sqrt{p(1-p)}}{p}$$

The number in each cell of Table 4.2.1 can be obtained by multiplying the proportion of failures (first expression) by the total number of individuals (second expression). Thus, for the first cell, which represents the number without the covariate in the control arm, there are $p(1 - r_2) * N(1 + r_1)/2$ individuals.

Table 4.2.1 Condensed 2x2x2 Table of Outcome, Treatment, and Covariate

	Without Covariate	With Covariate	Total
Control Arm	$p(1-r_2)$ $N(1+r_1)/2$	$p(1+r_2)$ $N(1-r_1)/2$	$p(1-r_1r_2)$ N
Treatment Arm	$p(1-r_2)$ $N(1-r_1)/2$	$p(1+r_2)$ $N(1+r_1)/2$	$p(1+r_1r_2)$ N
Total	$p(1-r_2)$ N	$p(1+r_2)$ N	p $2N$

From this, Canner developed an "influence" statistic for the distribution of covariate across outcome (Z_I) and a "disparity" statistic for the distribution of covariate across treatment arms (Z_D):

$$Z_I = \frac{p(1+r_2) - p(1-r_2)}{\sqrt{2pq/N}} = \frac{\sqrt{2N}pr_2}{\sqrt{pq}}$$

$$Z_D = \frac{\left[\left(\frac{N}{2}\right)(1+r_1) - \left(\frac{N}{2}\right)(1-r_1)\right]\sqrt{2N}}{\sqrt{N^2}} = \sqrt{2N}r_1$$

where $q = 1 - p$. The statistic for the unadjusted treatment effect (Z_U) is:

$$Z_U = \frac{p(1+r_1r_2) - p(1-r_1r_2)}{\sqrt{2pq/N}} = \frac{\sqrt{2N}pr_1r_2}{\sqrt{pq}}$$

and the adjusted treatment effect is equal to zero ($Z_A = 0$) by construction because the treatment effect is equal to zero. Thus, it follows that the difference

between the unadjusted and adjusted statistics is the product of the statistics for influence and disparity:

$$Z_U - Z_A = Z_U = \frac{\sqrt{2N}pr_1r_2}{\sqrt{pq}} = \frac{\left(\frac{\sqrt{2N}pr_2}{\sqrt{pq}}\right)\sqrt{2N}r_1}{\sqrt{2N}} = \frac{Z_I Z_D}{\sqrt{2N}}$$

This result can fully quantify the impact of failing to adjust for a covariate when there is no treatment effect. However, when a treatment effect is present, Z_A is no longer equal to zero. Beach and Meier extended Canner's method to present similar findings for a non-zero treatment effect in the context of linear regression parameters, which Canner[7] then re-derived in terms of correlations.

Canner presented the adjusted treatment effect as:

$$Z_A = \frac{b_{y1.2}}{\sqrt{\widehat{varb}_{y1.2}}} = \frac{r_{y1.2}\sqrt{N-3}}{\sqrt{1-r_{y1.2}^2}}$$

where $b_{y1.2}$ is the partial regression coefficient for outcome (Y) and treatment (X_1) given covariate (X_2) and $r_{y1.2}$ is the partial correlation coefficient for outcome (Y) and treatment (X_1) given covariate (X_2). It can also be shown that:

$$Z_A = \frac{(r_{y1} - r_{y2}r_{12}\sqrt{N-3})}{\sqrt{1-r_{y1}^2-r_{y2}^2-r_{12}^2+2r_{y1}r_{y2}r_{12}}}$$

given the well-known relationship:

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

where r_{y1} is the correlation between outcome (Y) and treatment (X_1), r_{y2} is the correlation between outcome (Y) and covariate (X_2), and r_{12} is the correlation between treatment (X_1) and covariate (X_2). This expression for Z_A can be used in conjunction with previous results to evaluate the impact of adjustment for a covariate when the treatment effect is non-zero.

4.3 Extension to Non-Inferiority Setting

The notation and results of Canner[6,7] and Beach and Meier[3] can be extended to develop an expression for the impact of adjustment for a covariate in the non-inferiority setting. In this scenario, the assumed treatment effect is non-zero and equal to the non-inferiority margin (d). The condensed 2x2x2 table can be written as:

Table 4.3.1 Condensed 2x2x2 Table of Outcome, Treatment, and Covariate with Treatment Effect Equal to Non-Inferiority Margin (d)

	Without Covariate	With Covariate	Total
Control Arm	$\frac{p(1-r_2)}{N(1+r_1)/2}$	$\frac{p(1+r_2)}{N(1-r_1)/2}$	$\frac{p(1-r_1r_2)}{N}$
Treatment Arm	$\frac{(p+d)(1-r_2)}{N(1-r_1)/2}$	$\frac{(p+d)(1+r_2)}{N(1+r_1)/2}$	$\frac{(p+d)(1+r_1r_2)}{N}$
Total	$\frac{(1/2)(2p+d(1-r_1))(1-r_2)}{N}$	$\frac{(1/2)(2p+d(1+r_1))(1+r_2)}{N}$	$\frac{(1/2)(2p+d(1+r_1r_2))}{2N}$

In this case, the new expression for the association between covariate and outcome is:

$$r_2^* = \frac{corr(cov, out)\sqrt{\frac{2p+d}{2}\left(1-\frac{2p+d}{2}\right)}}{\frac{2p+d}{2}}$$

From this, a new expression for influence (Z_{I^*}) is derived from the marginals as before:

$$Z_{I^*} = \frac{\left(\frac{1}{2}\right)\left(2p + d(1+r_1)\right)(1+r_2^*) - \left(\frac{1}{2}\right)\left(2p + d(1-r_1)\right)(1-r_2^*)}{\sqrt{\frac{pq + (p+d)\left(1-(p+d)\right)}{N}}}$$

$$Z_{I^*} = \frac{2pr_2^* + d(r_1 + r_2^*)}{\sqrt{\frac{pq + (p+d)(1-(p+d))}{N}}}$$

Canner's unadjusted statistic can then be re-written to include this:

$$Z_{C^*} = \frac{Z_{I^*}\, Z_D}{\sqrt{2N}}$$

and combined with the adjusted statistic as follows:

$$Z_{C^*} - Z_A = \frac{Z_{I^*}\, Z_D}{\sqrt{2N}} - \frac{\left(r_{y1} - r_{y2}r_{12}\right)\sqrt{2N-3}}{\sqrt{1 - r_{y1}^2 - r_{y2}^2 - r_{12}^2 + 2r_{y1}r_{y2}r_{12}}}$$

where the correlation between outcome and treatment (r_{y1}), the correlation between outcome and covariate (r_{y2}), and the correlation between treatment and covariate (r_{12}) are now equal to[28]:

$$r_{y1} = \frac{pr_1r_2^* + \left(\frac{1}{2}\right)d + \left(\frac{1}{2}\right)dr_1r_2^*}{\sqrt{p + \left(\frac{1}{2}\right)d + \left(\frac{1}{2}\right)dr_1r_2^* - p^2 - pd - pdr_1r_2^* - \left(\frac{1}{4}\right)d^2 - \left(\frac{1}{2}\right)d^2r_1r_2^* - \left(\frac{1}{4}\right)d^2r_1^2r_2^{*2}}}$$

$$r_{y2} = \frac{\left(\frac{1}{2}\right)dr_1 + \left(\frac{1}{2}\right)dr_2^* + pr_2^*}{\sqrt{p + \left(\frac{1}{2}\right)d + \left(\frac{1}{2}\right)dr_1r_2^* - p^2 - pd - pdr_1r_2^* - \left(\frac{1}{4}\right)d^2 - \left(\frac{1}{2}\right)d^2r_1r_2^* - (\frac{1}{4})d^2r_1^2r_2^{*2}}}$$

$$r_{12} = r_1$$

This new statistic $(Z_{C^*} - Z_A)$ can be used to evaluate the importance of adjustment for one covariate over another in a non-inferiority setting.

4.4 Relative Importance of r_1 and r_2

To evaluate the impact of varying scenarios on this statistic for importance $(Z_{C^*} - Z_A)$ in a non-inferiority setting, a simulation study was undertaken. The simulation explored this statistic's behavior when subject to changes in r_1 and r_2 for given proportions of failure (or success) (p) and non-inferiority margins (d). Figure 4.4.1 illustrates the impact of r_1 and r_2 when p is 0.4, d is 0.1, and r_1 and r_2 are varied from -0.9 to 0.9. In this scenario, N is set to be 388 per arm to ensure 80% power.

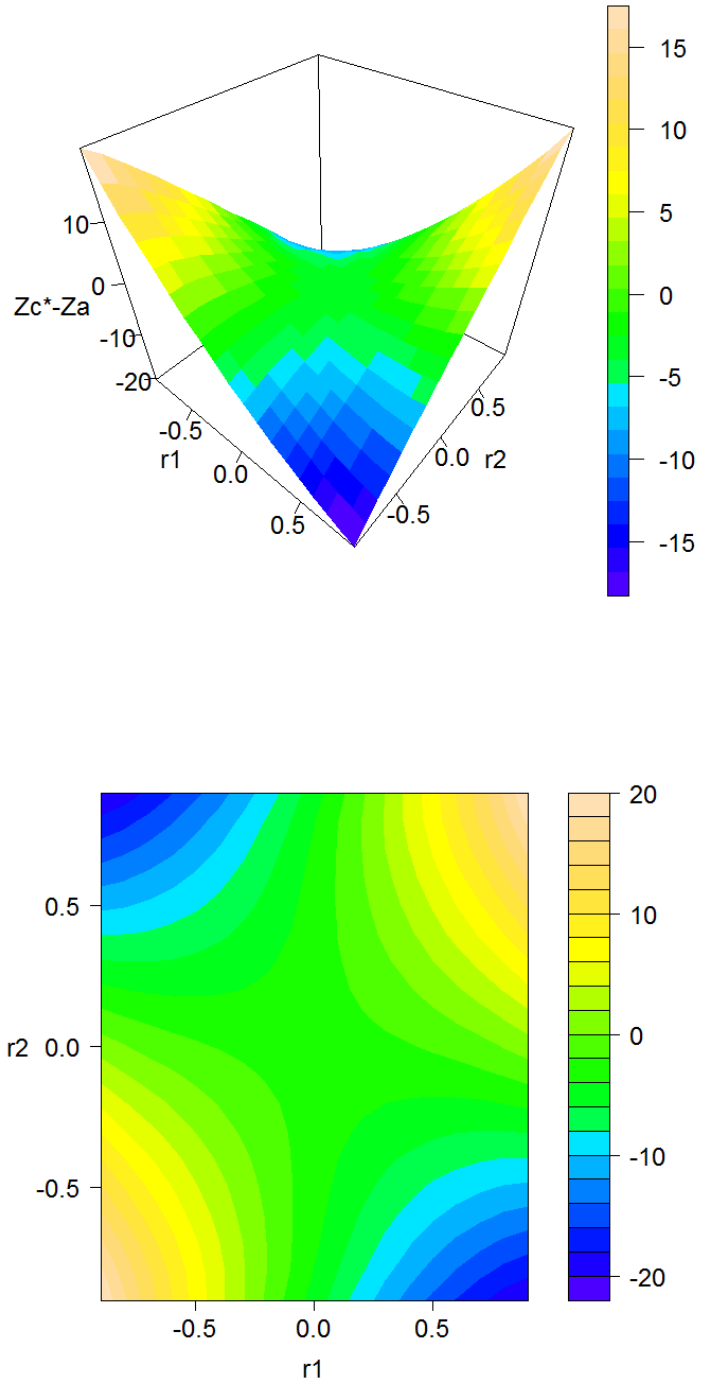

Figure 4.4.1 Z_{C^*}-Z_A for Full Range of r_1 and r_2 when $N = 388, p = 0.4, d = 0.1$

The statistic for the importance of adjustment is presented as the contour and ranges from -20.35 to 20.30. The green represents values of the statistic close to $-Z_A$, which occur when either $r_1$or $r_2$are close to zero. The point $(r_1, r_2) = (0,0)$ occurs at the center of both the three dimensional and contour plots. The yellow and beige represent increasingly positive values of the statistic, which occur when r_1 and r_2 are in the same direction (i.e., both positive or both negative); blue and purple represent increasingly negative values of the statistic, which occur when r_1 and r_2 are in opposite directions (i.e., one positive and one negative). Furthermore, a near symmetry exists such that similar values of the statistic are obtained when r_1 is positive and r_2 is negative as when r_1 is negative and r_2 is positive, and the same is true for associations in the same direction.

We can see that according to our statistic, evaluation of either imbalance (r_1) or strength of association (r_2) in isolation is not sufficient to identify the most important subset of covariates. Rather, adjustment is necessary when a covariate is both imbalanced and strongly associated with outcome, and these two forces are given similar weight. We will now narrow our focus to more practical values (i.e., where r_1 and r_2 range from -0.2 to 0.2) to explore the relative importance of these two factors.

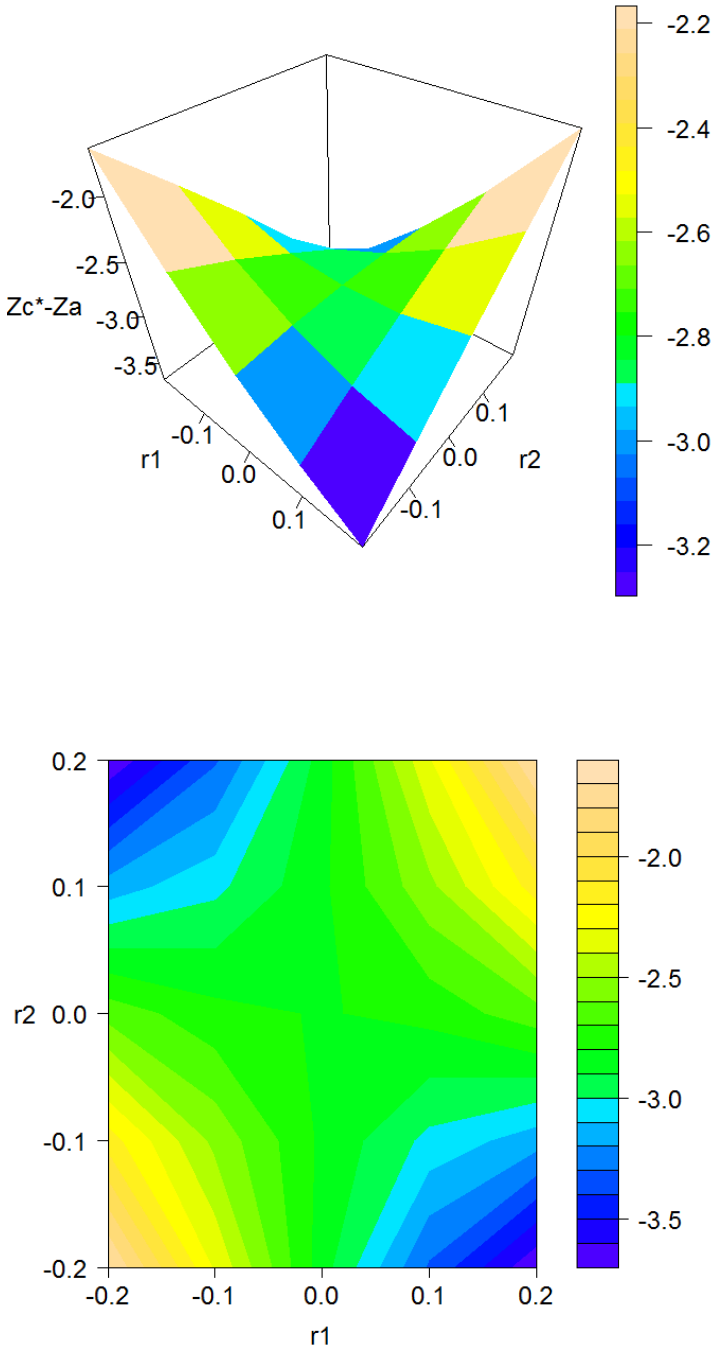

Figure 4.4.2 Z_{C^*}-Z_A for Reduced Range of r_1 and r_2 when $N = 388, p = 0.4, d = 0.1$

Given this reduced range, it becomes apparent that even for moderate strengths of association (i.e., r_1 and r_2 between -0.2 and 0.2) the statistic is close to $-Z_A$ and adjustment is not necessary. However, when either r_1 or r_2 is large and the other is non-negligible, adjustment is necessary. Such findings are consistent with previous research on covariate adjustment in the presence of imbalance.[1,3,8,29,31,36]

In addition, it is important to note that while the magnitude of the statistic is dependent on the sample size and probability of failure, the relative importance given to r_1 and r_2 remains generally consistent. For example, Figure 4.4.3 shows the statistic when p is 0.1 (N is set to be 199 to maintain the same power as above). With the probability of failure in the standard of care now equal to the treatment effect (i.e., the non-inferiority margin), the impact of the treatment effect is more apparent. The range of the statistic is reduced and, while the shape remains similar, the asymmetry is more pronounced.

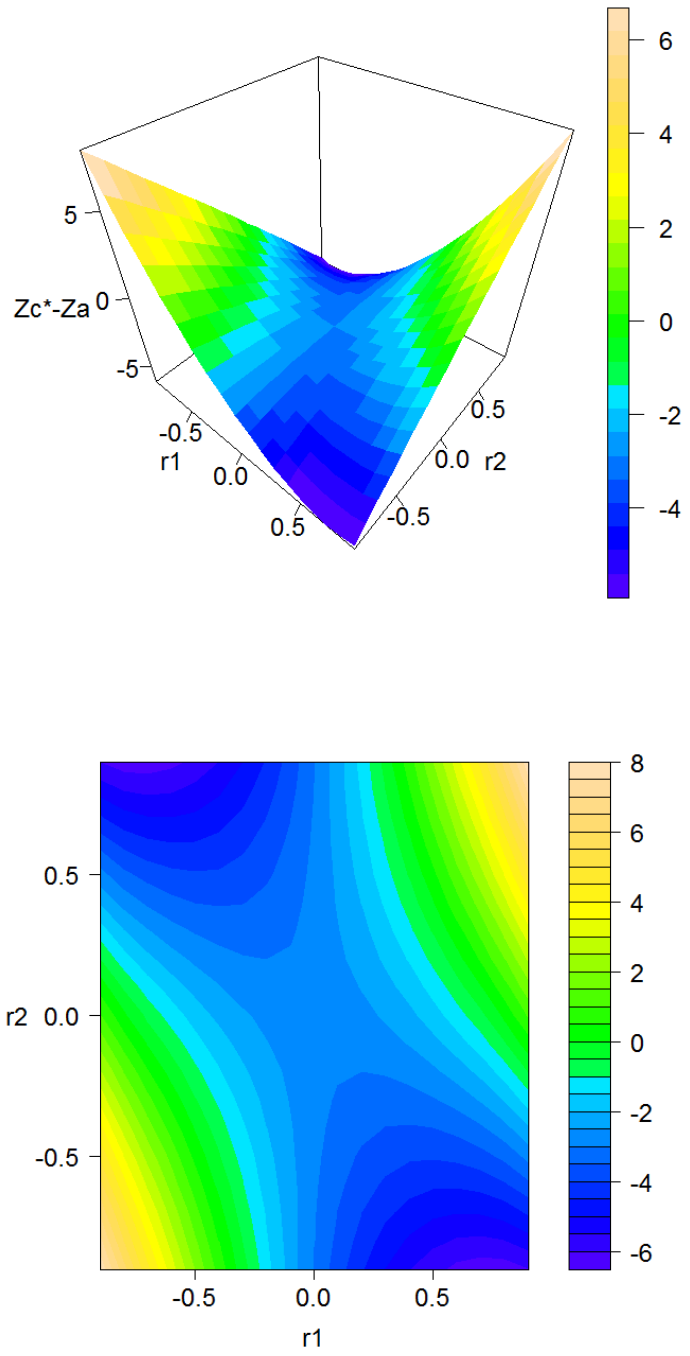

Figure 4.4.3 Z_{C^*}-Z_A for Full Range of r_1 and r_2 when $N = 199, p = 0.1, d = 0.1$

4.5 Relationship to Bias

The proposed statistic also has an important relationship to the bias in the estimate of the treatment effect on the risk difference scale. The bias in the original case where p is 0.4, d is 0.1, and N is 388 is presented below as the unadjusted estimate minus the adjusted estimate $(\beta_{unadj} - \beta_{adj})$.

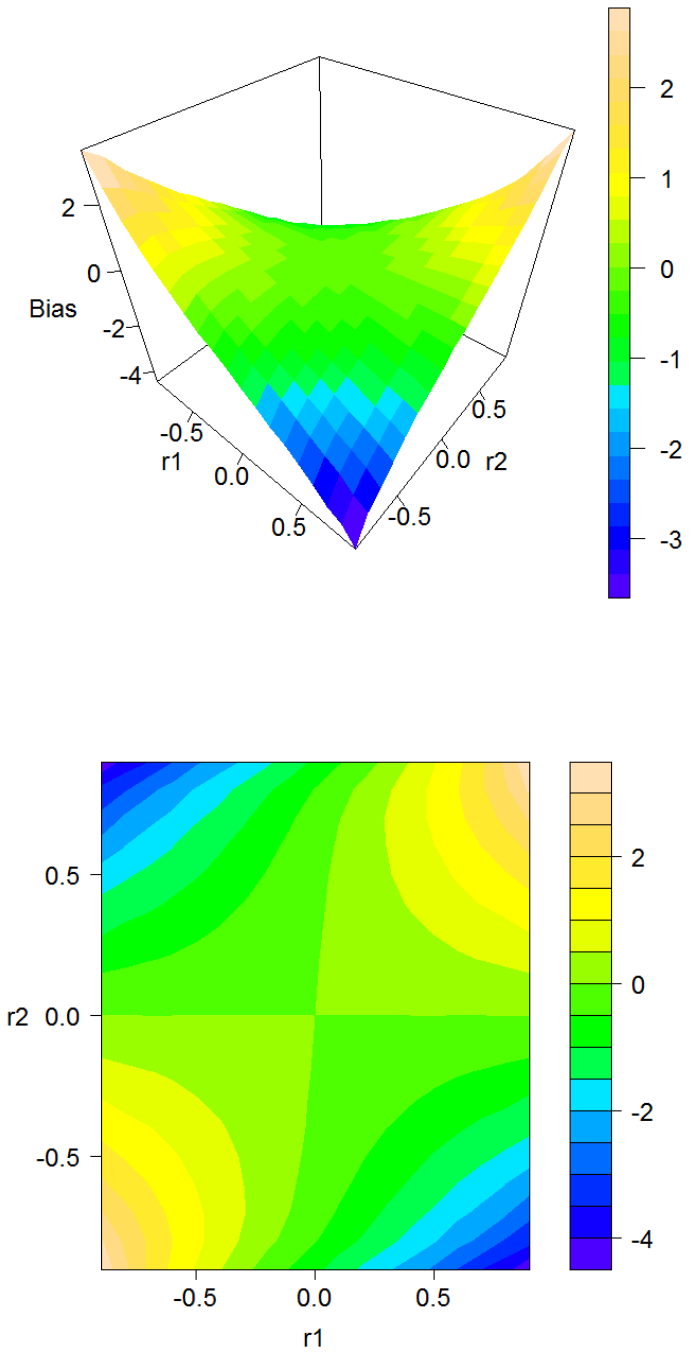

Figure 4.5.1 Bias for Full Range of r_1 and r_2 when $N = 388, p = 0.4, d = 0.1$

From the comparison of Figure 4.5.1 to Figure 4.4.1, one can see that the statistic for importance of adjustment maps nicely to the bias in the treatment estimate. Given r_1 and r_2 equal to zero, the statistic is equal to $-Z_A$ and the estimate of the treatment effect is unbiased. When r_1 and r_2 are in the same direction (i.e., both positive or both negative), the unadjusted estimate is biased upward, and when r_1 and r_2 oppose each other (i.e., one is positive and the other is negative), the unadjusted estimate is biased downward. In the case of a two-sided test, this is an unimportant distinction. However, in the context of non-inferiority, a one-sided test (or confidence interval) is used, and this directionality becomes important.

4.6 Implications for Error

Adjustment for covariates will always slightly increase the width of the confidence interval of the treatment effect in the context of logistic regression, and the degree to which it changes is primarily dependent on the magnitude (but not the directionality) of r_1.[33] However, as shown in Figure 4.6.1, this increase only becomes important at extreme values of r_1 and r_2.

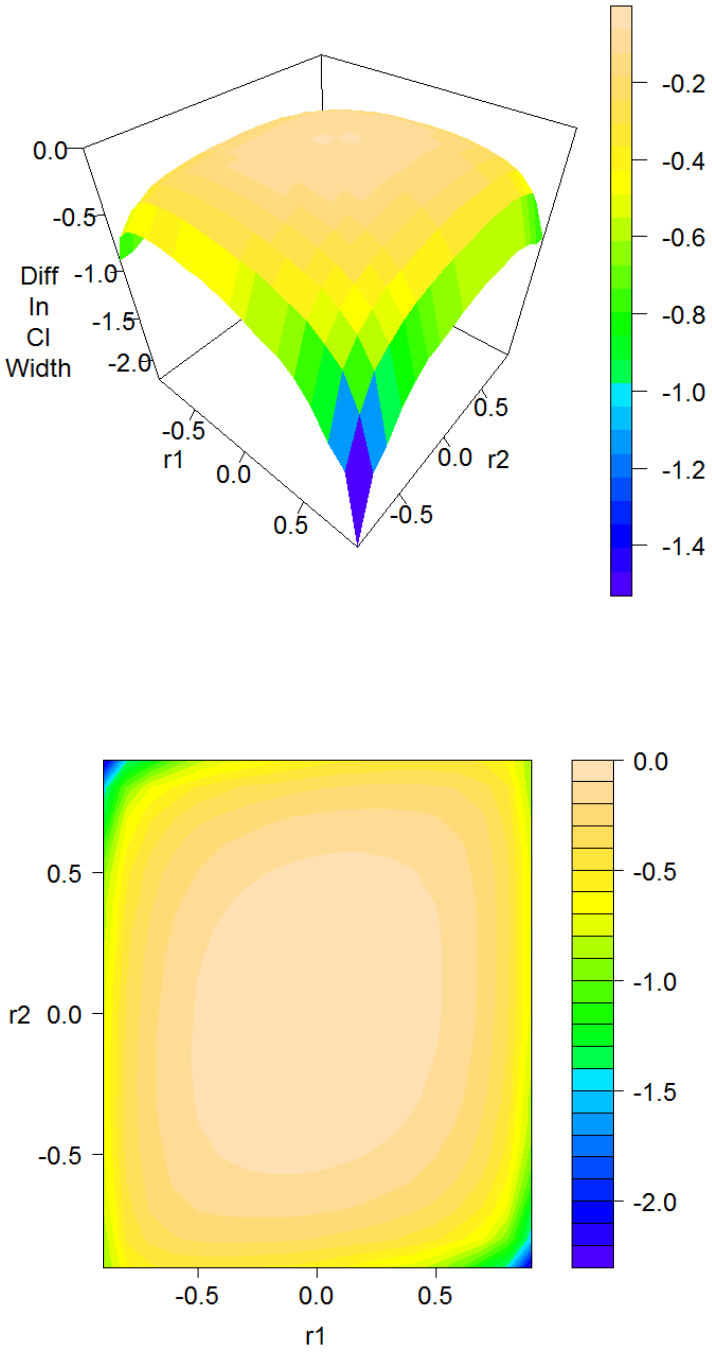

Figure 4.6.1 Change in Width of the 95% Confidence Interval (Unadjusted – Adjusted) for Full Range of r_1 and r_2 when $N = 388, p = 0.4, d = 0.1$

If one is concerned about whether the non-inferiority margin is crossed, as in the case of a confidence interval approach to a non-inferiority hypothesis, both the point estimate and its variance must be considered. As in the theoretical schematic presented in Figure 4.6.2, while the increased variance imposed by adjustment may push the upper bound beyond the margin, decreasing the potential to reject the null hypothesis of inferiority, this impact is small in comparison to the change in the estimate itself. Thus, there is a serious potential for committing a type I error when failing to adjust for a covariate that has imbalance and influence in the same direction, regardless of whether that direction is positive or negative.

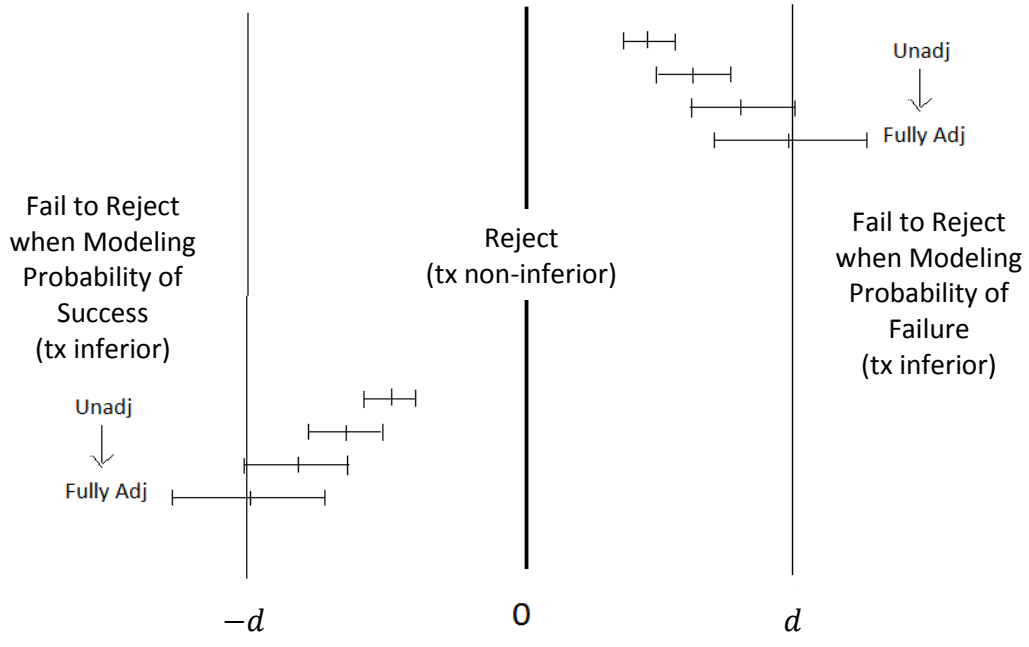

Figure 4.6.2 Movement of the 95% Confidence Interval Given Adjustment and Implications for Error Probabilities in Non-Inferiority Trials

4.7 Practical Use

In practice, the statistic for importance of adjustment can be used as a diagnostic tool. To this end, one might calculate values of the statistic for all available covariates and then rank them. The covariate corresponding to the largest value would be the most important covariate for which to adjust and so on. For illustrative purposes, consider a plot of covariates corresponding to all possible combinations for $r_1$and r_2 with N, p, and d again equal to 388, 0.4, and 0.1, respectively.

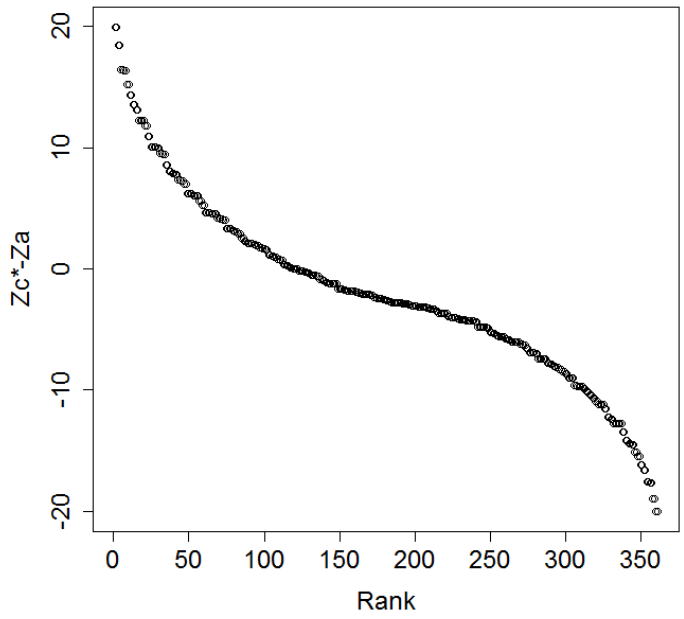

Figure 4.7.1: Ranking of Covariates with All Possible Combinations of Full Range of $r_1$and r_2 where $N = 388$, $p = 0.4$, and $d = 0.1$

In a real trial setting, it is likely that there will be less than ten covariates under consideration. Thus, random subsets are presented:

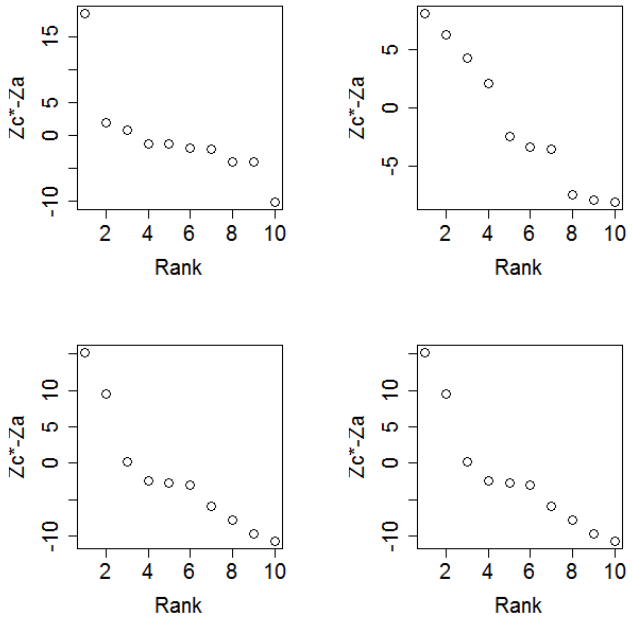

Figure 4.7.2: Random Subsets of Ten Covariates from Figure 4.7.1

Generally, one would expect to see an "elbow" or plateau which could be used as the cut point for adjustment. In so doing, one can choose to include only those covariates falling above the cut point in the final analysis model. Such simultaneous consideration of both influence and disparity allows a trialist to find the optimal balance between the competing goals of modeling a treatment effect that is both precise and conservative.

4.8. Discussion

It has been shown that failure to adjust for covariates in a non-inferiority trial with binary outcomes leads to biased estimates of the treatment effect and increased type I error.[18,27] As such, the issue becomes not whether one should adjust, but rather which covariates should be included. While many model selection procedures are theoretically akin to the statistic for influence and tests of baseline imbalance are theoretically akin to the statistic for disparity, neither approach considers these effects simultaneously. The proposed statistic allows for evaluation of the importance of adjustment for each covariate relative to the others and, when used properly, should result in a better estimate of treatment effect in terms conservativeness and precision given the available data.

The authors recognize that, like tests of baseline imbalance and other data-driven approaches to covariate adjustment, the use of this statistic may have important implications for type I error that should be further quantified in a multivariate setting. Berger[4] has proposed an extension whereby the disparity of a candidate covariate is calculated as the weighted average of the disparities within each group formed by the covariates already present in the model. The practical use of this approach, however, is difficult to explore. The lack of information present in binary variables leads to highly variable results such that iterative methods may not be sufficient to assure consistency. Thus, while this has been evaluated in the context of linear regression[3], it remains largely unexplored

in the context of binary outcomes. Still, the proposed statistic remains useful as a diagnostic tool when used in conjunction with other techniques for covariate selection.

Chapter 5

Aim 3:

Application of a Joint Statistic for Influence and Disparity in the Identification of Baseline Covariates Required In Analysis of Non-Inferiority Trials

Due to the increase in type I error imposed by exclusion of important prognostic variables in non-inferiority trials, a joint statistic for the importance of adjusting for a given covariate based on its association with outcome (influence) and its association with treatment (imbalance) has been developed for use when important factors are not known *a priori*. Once the data has been collected, this statistic can be used to rank the relative importance of adjusting for one covariate over another, and its distribution can be used to determine the optimal number of covariates to include in analysis. The goal of the present work is to evaluate the practicality of this method of covariate selection in the rapid anti-convulsant medication prior to arrival trial (RAMPART) in terms of changes to treatment estimates and possible implications for maintenance of type I error.

5.1 Introduction

The ICH guidelines[22] for the use of baseline covariates in clinical trials states:

> "Pre-trial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis in order to improve precision and to compensate for any lack of balance between treatment groups. If one or more factors are used to stratify the design, it is appropriate to account for those factors in the analysis. When the potential value of an adjustment is in doubt, it is often advisable to nominate the unadjusted analysis as the one for primary attention, the adjusted analysis being supportive."

Thus, according to this guidance, covariates should be specified *a priori* and those that are deemed necessary for inclusion should be present in both the randomization and the analysis. However, potential prognostic factors are not always known at the start of a trial. Although adjusted analyses offer more power to detect a treatment difference in superiority trials[15-17,26,33], unadjusted analyses may not necessarily inflate type I error. This does not hold for non-inferiority designs.

In non-inferiority trials, unadjusted analyses can be plagued by sometimes dramatic increases in type I error in the presence of important prognostic variables. According to the author's previous work[27], these rates can be doubled depending on the sample size, the probability of success in the standard of care group, the underlying treatment difference, and the covariate's strength of association with outcome. Thus, it is necessary to develop an approach to identify

important prognostic covariates in non-inferiority trials where such factors are not known *a priori*. Although much of the statistical literature condemns the use of data-driven approaches in clinical trials due to the increased probability of selection bias,[1,4.31,32,35,36] such work has been done in the context of the superiority design and has not been fully investigated in non-inferiority trials.

The two factors contributing to the impact of a covariate on the estimate of the treatment effect are its association with treatment assignment (i.e., imbalance) and its association with outcome. If precision is the goal, then a trialist should adjust for those baseline covariates that are imbalanced across treatment groups, but if accuracy is the goal, then those covariates that are associated with outcome, regardless of imbalance, should be included.[3,7,32] As Canner[7] has illustrated, to obtain a treatment estimate that is both precise and accurate, these factors must be considered in tandem. Thus, a joint statistic has been developed to evaluate these aspects simultaneously for a given covariate in a trial with dichotomous outcome and has been subsequently extended for use in the non-inferiority setting.[28] The aim of the current research is to evaluate the use of this joint statistic in a practical clinical trial application.

5.2 Existing Literature

It has been shown that failure to adjust for covariates in a non-inferiority trial with a non-linear outcome leads to biased estimates of the treatment effect

and increased type I error.[18,27] As such, the issue becomes not whether one should adjust, but rather which covariates should be considered for inclusion in the analysis. While many model selection procedures are theoretically akin to the statistic for influence and tests of baseline imbalance are theoretically akin to the statistic for disparity, neither approach considers these effects simultaneously. Canner[6] developed a statistic to quantify a covariate's disparity and influence given a binary outcome and zero treatment effect, which Beach and Meier[3] extended for non-zero treatment effects with continuous outcome. In the author's previous paper[28], Canner and Beach and Meier's work were used to develop a statistic for the non-inferiority framework with binary outcome and non-null treatment effect. This statistic, which represents the product of statistics for a covariate's influence on outcome (Z_I^*) and disparity across treatment groups (Z_D), can be expressed as follows:

$$Z_{C^*} - Z_A = \frac{Z_{I^*}\, Z_D}{\sqrt{2N}} - \frac{\left(r_{y1} - r_{y2}r_{12}\right)\sqrt{2N-3}\right)}{\sqrt{1 - r_{y1}^2 - r_{y2}^2 - r_{12}^2 + 2r_{y1}r_{y2}r_{12}}}$$

where

$$Z_{I^*} = \frac{2pr_2^* + d(r_1 + r_2^*)}{\sqrt{\frac{pq + (p+d)(1-(p+d))}{N}}}$$

$$Z_D = \frac{\left[\left(\frac{N}{2}\right)(1+r_1) - \left(\frac{N}{2}\right)(1-r_1)\right]\sqrt{2N}}{\sqrt{N^2}} = \sqrt{2N}r_1$$

$$r_1 = r_{12} = corr(trt, cov)$$

$$r_2^* = \frac{corr(cov, out)\sqrt{\frac{2p+d}{2}\left(1-\frac{2p+d}{2}\right)}}{\frac{2p+d}{2}}$$

$$r_{y1} = \frac{pr_1r_2^* + \left(\frac{1}{2}\right)d + \left(\frac{1}{2}\right)dr_1r_2^*}{\sqrt{p + \left(\frac{1}{2}\right)d + \left(\frac{1}{2}\right)dr_1r_2^* - p^2 - pd - pdr_1r_2^* - \left(\frac{1}{4}\right)d^2 - \left(\frac{1}{2}\right)d^2r_1r_2^* - \left(\frac{1}{4}\right)d^2r_1^2r_2^{*2}}}$$

$$r_{y2} = \frac{\left(\frac{1}{2}\right)dr_1 + \left(\frac{1}{2}\right)dr_2^* + pr_2^*}{\sqrt{p + \left(\frac{1}{2}\right)d + \left(\frac{1}{2}\right)dr_1r_2^* - p^2 - pd - pdr_1r_2^* - \left(\frac{1}{4}\right)d^2 - \left(\frac{1}{2}\right)d^2r_1r_2^* - (\frac{1}{4})d^2r_1^2r_2^{*2}}}$$

$$N = sample\ size\ per\ arm$$

$$p = probability\ of\ failure$$

$$d = treatment\ effect$$

By calculating this statistic for all available covariates in the study data and ranking, a trialist can evaluate the importance of adjusting for one covariate over another in a non-inferiority setting. The covariate ranked highest is the most important one to adjust for and each subsequent covariate becomes less so.[28] The top ranks can then be used in a model building procedure to determine the most appropriate multivariate model in terms of both the precision of the estimate of the treatment effect and the extent to which it maintains the type I error.

5.3 Application

5.3a RAMPART

The Rapid Anti-Convulsant Medication Prior to Arrival Trial (RAMPART) was a randomized double dummy phase III clinical trial designed to test whether intramuscular (IM) midazolam was non-inferior by a margin of 10% to intravascular (IV) lorazepam for the treatment of status epilepticus, operationally defined as a seizure lasting at least five minutes without interruption. Because the disease is life-threatening, the paramedics needed to administer the treatment immediately upon arrival. Thus, this study was conducted in an ambulatory setting under exception from informed consent, and the randomization scheme did not allow for any covariate adjustments.[40] Subjects were well balanced between treatment groups on demographic and clinical characteristics, and the primary outcome of seizure termination prior to ED arrival without the use of rescue medication was met for 329 of 448 (73.4%) subjects allocated to active IM treatment and in 282 of 445 (63.4%) allocated to active IV treatment (risk difference (RD): 10.1%, 95% CI: 4.0%,16.1%). Thus, the trial results showed that IM midazolam was actually superior to IV lorazepam.[39] The primary analysis did not adjust for any pre-specified prognostic variables. However, this secondary analysis explores whether or not the inclusion of some subset of available covariates may improve the precision and conservativeness of the treatment effect.

The statistic for importance was calculated for several potential covariates of interest including subsequent ICU admission, hospital admission, intubation, recurrent seizure within 12 hours, hypotension, Hispanic ethnicity, type of seizure, prior history of seizure, low dose vs high dose, non-white race, pediatric vs adult, intubation within 30 minutes, recurrent seizure, and gender. Table 5.3a.1 gives the various components of the statistic for importance, including r_1 (the correlation between treatment and covariate), r_2 (a function of the correlations between covariate and outcome and treatment and outcome), Z_D (the statistic for disparity), and Z_I^* (the statistic for influence), as well as the estimate of the treatment effect on the risk difference scale, including the associated 95% confidence interval and bias.

Table 5.3a.1: Statistics and Treatment Estimates by Covariate for RAMPART. Each Model Contained Treatment Plus One Covariate.

	r_1	r_2	Z_D	Z_I^*	Z Importance	Lower Bound*	Treatment Estimate	Upper Bound*	Bias in the Treatment Estimate**
ICU Admission	-0.0813	-0.1321	-2.4298	-6.1057	-2.3383	0.0267	0.0866	0.1464	0.0141
Hospital Admission	-0.0825	-0.1222	-2.4664	-5.6757	-2.3880	0.0274	0.0875	0.1475	0.0132
Intubation	-0.0470	-0.1383	-1.4033	-6.2676	-2.7328	0.0324	0.0920	0.1515	0.0087
Recurrent Seizure (12 Hrs)	0.0132	0.0783	0.3930	3.5059	-3.1794	0.0390	0.0993	0.1596	0.0014
Hypotension	-0.0074	-0.0507	-0.2199	-2.2651	-3.2254	0.0396	0.1002	0.1607	0.0005
Hispanic	-0.0289	-0.0077	-0.8646	-0.4359	-3.2285	0.0397	0.1004	0.1612	0.0002
Type of Seizure	0.0086	0.0195	0.2559	0.8888	-3.2359	0.0397	0.1004	0.1611	0.0002
Prior Hx	-0.0094	-0.0045	-0.2805	-0.2315	-3.2456	0.0399	0.1006	0.1613	0.0000
Low Dose	0.0085	0.0010	0.2536	0.0722	-3.2486	0.0399	0.1007	0.1614	0.0000
Non-White	0.0007	-0.0003	0.0216	-0.0130	-3.2495	0.0399	0.1007	0.1614	0.0000
<18 Yrs of Age	-0.0013	0.0091	-0.0396	0.3980	-3.2509	0.0400	0.1007	0.1614	0.0000
Intubation(30 Mins)	-0.0046	-0.1308	-0.1367	-5.7985	-3.2584	0.0403	0.0999	0.1594	0.0008
Recurrent Seizure	-0.0013	0.0611	-0.0380	2.7003	-3.2698	0.0403	0.1008	0.1612	-0.0001
Gender	-0.0233	0.0194	-0.6964	0.7816	-3.2902	0.0406	0.1013	0.1620	-0.0007

*For 95% confidence interval of the treatment effect

**Bias is calculated as $\beta_{unadj} - \beta_{adj}$

The table is currently sorted by Z Importance, which is a weighted product of Z_D and Z_I^* minus the adjusted Z value. Thus, it represents a joint measure of disparity and influence. By ranking the covariates in this manner, we are able to not only identify the most important subset in terms of both the point estimate itself and its precision, but also to consider covariates relative to each other. Other rankings can be explored. For example, if one were concerned with precision, one might choose to rank by Z_D. In this case, hospital admission would surpass ICU admission as the top ranked covariate. However, if the goal was accuracy, one might chose to use Z_I^* instead, revealing intubation, ICU admission, and intubation within 30 minutes. Intubation within 30 minutes is now among our top three because it has a large r_2 relative to the others. However, if we are interested in both the point estimate and its precision, we would not want to choose intubation within 30 minutes as its r_1 is quite small. Figure 5.3a.1 illustrates the impact on the treatment estimate on the risk difference scale when a specific covariate is included in the model.

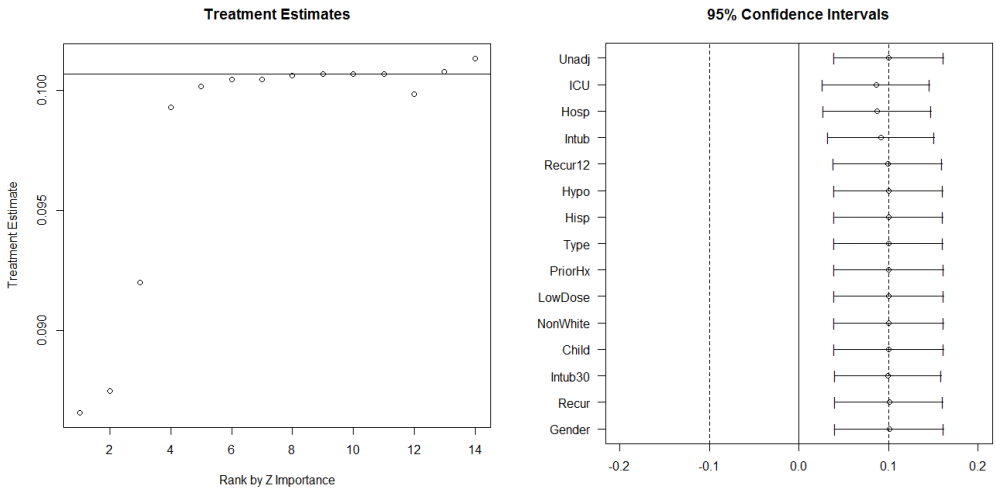

Figure 5.3a.1: Treatment Estimates and 95% Confidence Intervals for Covariates in RAMPART. Each Model Contains Treatment Plus One Covariate.

Those covariates with high values of r_1 and r_2 in the same direction pull the point estimate of the treatment effect down (toward the null hypothesis) regardless of whether that direction is postive or negative, whereas those covariates with correlations in oppostie directions (i.e., one postive and one negative) pull the estimate up. Thus, when controlling for type I error is the goal, as it should be given the increased potential under the non-inferioirty design, those covariates with correlation not in opposition are ranked highest. For the top ranked covariates, both the treatment estimate and the variance are reduced and this reduction is less pronounced as we move through the ranks. Figure 5.3a.2 shows the values of Z importance for the covariates.

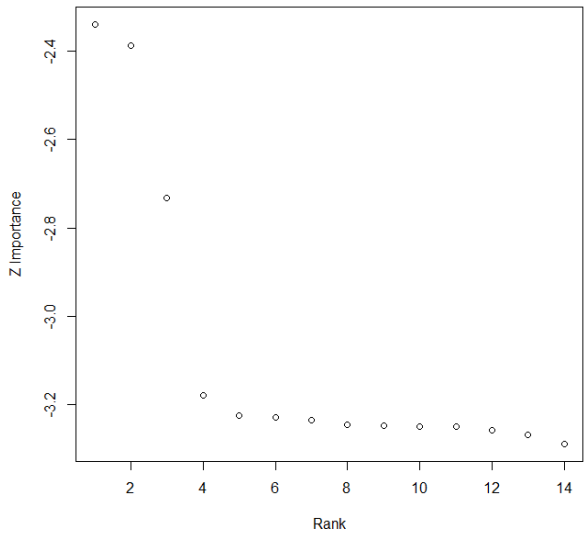

Figure 5.3a.2: Z Importance for Covariates in RAMPART

The same information that is presented in Figure 5.3a.1 is quantified in Z Importance. From the "elbow" in the curve, it is clear that two, or perhaps three, covariates are worth further evaluation. These are ICU admission, hospital admission, and intubation. For the analysis model, ICU admission and hospital admission cannot both be included in the analysis because of the high probability of co-linearity, so we can choose the higher ranked of these and conduct the analysis adjusting for ICU admission and intubation. Figure 5.3a.3 illustrates the unadjusted and adjusted confidence intervals for the treatment effect.

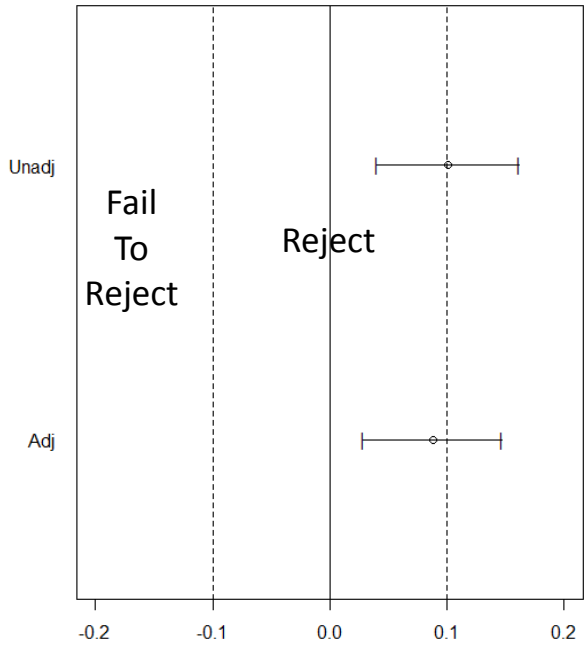

Figure 5.3a.3: Impact of Adjustment Using Z Importance in RAMPART

Although we will never reach the non-inferiority bound set for this trial (RD = -10%) due to the inherent superiority of IM Midazolam that was confirmed in the trial, we can see that the adjusted analysis shifts the point estimate toward the null hypothesis in this setting, protecting against the probability of committing a type I error. Furthermore, we have discovered a more precise estimate of the treatment effect in this study.

5.3b RAMPART Pediatric Population

As the RAMPART study represents the largest prospective clinical trial experience with prehospital status epilepticus (SE) in children and adolescents, a secondary analysis was performed to describe and clinically characterize this unique cohort. Of the 893 subjects in the primary intent to treat analysis, 120 were less than 18 years of age. As in the primary analysis, the pediatric population was well balanced across treatment groups in terms of demographic and clinical characteristics, and the primary outcome of seizure termination prior to ED arrival without the use of rescue medication was met for 41 of 60 (68.3%) patients randomized to active IM treatment and 43 of 60 (71.7%) randomized to active IV (RD: -3.3%, 99% CI=-24.9%, 18.2%). Although the study was not powered for a secondary analysis of the pediatric population, IM Midazolam did not indicate superiority over IV Lorazepam as it did in the primary analysis. However, it did show non-inferiority in that the 99% confidence interval included the 10% non-inferiority margin. Furthermore, there were some important differences favoring IM Midazolam in the pediatric population that were not evident in the cohort as a whole including lower rates of recurrent seizure, intubation, and ICU admittance.[43] Thus, additional analyses that account for these covariates may reveal important findings in the cohort in addition to improving the accuracy and precision of the treatment estimate.

The statistic for the importance of adjustment was calculated for the same covariates considered above with the exception of hypotension as none of subjects in the pediatric population met this criterion. Table 5.3b.1 provides the values of Z Importance, as well as its components, and the treatment estimates.

Table 5.3b.1: Statistics and Treatment Estimates by Covariate for Pediatric Sub-Population of RAMPART. Each Model Contained Treatment Plus One Covariate.

	r_1	r_2	Z_D	Z_I^*	Z Importance	Lower Bound*	Treatment Estimate	Upper Bound*	Bias in the Treatment Estimate**
Intubation	-0.1038	-0.1409	-1.1375	-2.3182	0.8959	-0.2172	-0.0544	0.1084	0.0211
Hispanic	-0.0933	-0.1360	-1.0222	-2.2396	0.8292	-0.2144	-0.0515	0.1113	0.0182
ICU Admission	-0.1097	-0.0846	-1.2013	-1.3727	0.7059	-0.2123	-0.0469	0.1186	0.0136
Hospital Admission	-0.0834	-0.0691	-0.9140	-1.1243	0.5875	-0.2072	-0.0417	0.1238	0.0084
Intubation (30 Mins)	-0.0949	-0.0588	-1.0398	-0.9460	0.5800	-0.2075	-0.0415	0.1245	0.0082
Gender	0.0170	0.0896	0.1857	1.4938	0.4482	-0.1999	-0.0355	0.1289	0.0021
Type of Seizure	-0.0268	-0.0345	-0.2937	-0.5665	0.4249	-0.2004	-0.0347	0.1311	0.0013
Low Dose	-0.0219	-0.0407	-0.2403	-0.6733	0.4241	-0.2003	-0.0346	0.1311	0.0013
Prior Hx	-0.0338	0.0000	-0.3703	0.0135	0.3934	-0.1994	-0.0334	0.1327	0.0000
Non-White	-0.0336	0.1345	-0.3684	2.2661	0.2497	-0.1896	-0.0270	0.1355	-0.0063
Recurrent Seizure	-0.1206	0.0431	-1.3212	0.7693	0.2178	-0.1933	-0.0264	0.1404	-0.0069
Recurrent Seizure (12 Hrs)	-0.1336	0.0668	-1.4639	1.1720	0.0928	-0.1879	-0.0212	0.1454	-0.0121

*For 95% confidence interval of the treatment effect

**Bias is calculated as $\beta_{unadj} - \beta_{adj}$

The covariates in Table 5.3b.1 are ranked by Z Importance, which represents a weighted product of the Z for disparity (Z_D) and the Z for influence (Z_I^*) minus that adjusted Z value. If one were to evaluate r_1 (or Z_D) in isolation, recurrent seizure (and the further refined recurrent seizure within 12 hours) would show the greatest imbalance across treatment arms. However, imbalance alone is not sufficient to justify inclusion of covariates in analysis. In fact, as described previously, covariates must be both disparate and influential to have important implications for the treatment effect, and the r_2 (or Z_I^*) for these are quite low in comparison to the other available covariates. Similarly, non-white race shows high influence, but low disparity. Figure 5.3b.1 illustrates this impact for all available covariates in terms of point estimates and confidence intervals.

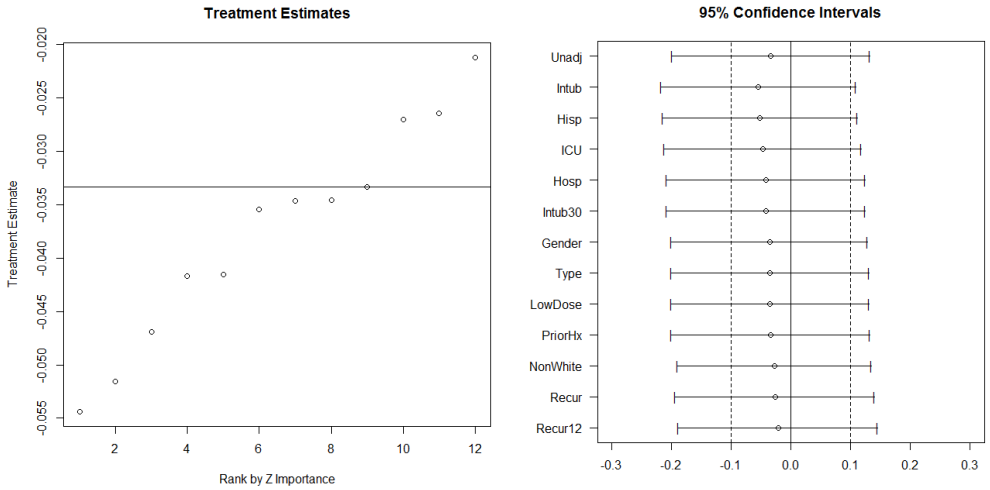

Figure 5.3b.1 Treatment Estimates and 95% Confidence Intervals for Covariates in Pediatric Sub-Population of RAMPART. Each Model Contains Treatment Plus One Covariate.

In the pediatric cohort, as with the population as a whole, the covariates with correlation in the same direction pull the point estimate of the treatment effect down (towards the null hypothesis), while those with opposing correlations pull it up. It is those covariates that pull it down which are important to include in order to maintain type I error. However, the relative impact on the point estimate and its variance diminishes as we continue to add covariates to the model. Figure 5.3b.2 provides the distribution of the values of Z Importance, which can be used to determine the best number of covariates for inclusion.

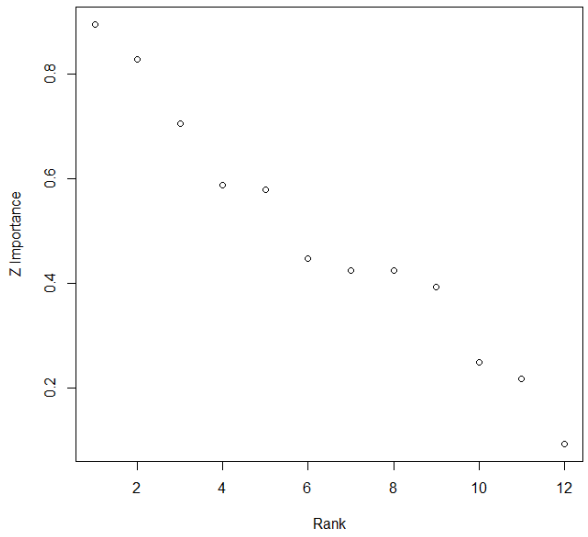

Figure 5.3b.2: Z Importance for Covariates in Pediatric Sub-Population of RAMPART

The distribution in Figure 5.3b.2 does not show the same plateau effect as in the combined population of children and adults (Figure 5.3a.2), making the decision

less obvious. Thus, one might chose to evaluate several models for sensitivity. In this case, however, the issue of collinearity resurfaces. The top ranked covariates, in order of importance, are intubation, Hispanic ethnicity, ICU admission, hospital admission, and intubation within 30 minutes of ED arrival. The top three are likely to require inclusion, while hospital admission and intubation within 30 minutes are more debatable. As before, in the event of likely collinearity, it is reasonable to take the higher ranked covariate in each case. Thus, the most conservative and precise estimate of treatment effect results from adjustment for intubation, Hispanic ethnicity, and ICU admission. Figure 5.3b.3 provides this estimate and its 95% confidence interval along with the unadjusted estimate for comparison.

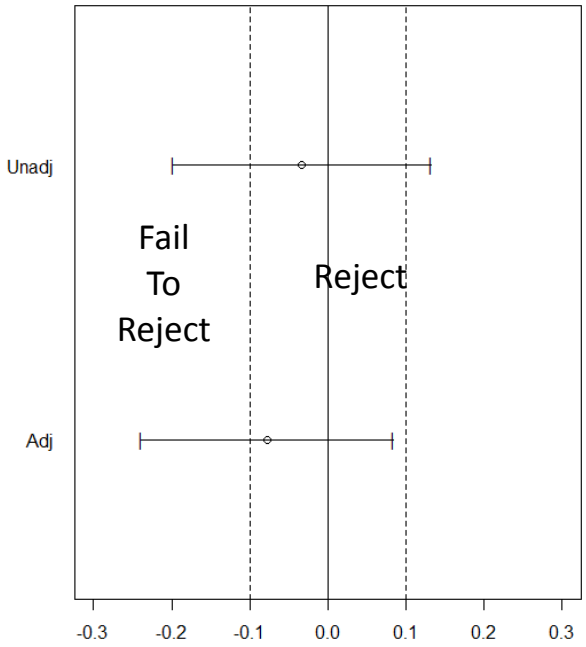

Figure 5.3b.3: Impact of Adjustment Using Z Importance in Pediatric Sub-Population of RAMPART

The inclusion of intubation, Hispanic ethnicity, and ICU admission provides a more precise estimate of the treatment effect, and, although the overall finding of the trial has not been changed, the use of the joint statistic for disparity and influence has revealed a more conservative model.

5.4 Discussion

Failure to adjust for important covariates in non-inferiority trials can have serious implications for type I error (claiming a treatment is non-inferior when it

truly is inferior).[18.27] As such, we need to make special efforts to protect against type I errors when operating under this framework. A statistic for the importance of including a given covariate in analysis has been previously developed by the authors and combines information about a covariate's imbalance across treatment groups and influence on outcome. This statistic can be calculated for all available covariates in a trial and used to rank them. Based on the distribution of these values, an appropriate number of covariates can be selected for inclusion in analysis. The end result will be an estimate of treatment effect that more precise. Furthermore, the inclusion of variables based on this method will ensure conservative estimates in terms of type I error under a non-inferiority study design.[28]

When applied to a real trial dataset, the statistic was able to identify those variables that maximized the tradeoff between changes in the point estimate itself and its precision and produced conservative estimates of treatment effect. Although the inclusion of the covariates in analysis did not change the overall conclusions of either trial, both resulted in a more conservative measure of the assumed treatment effect. One limitation of this work is that some of the covariates under consideration, such as intubation, recurrent seizure, and hospital or ICU admission, may not necessarily constitute prognostic factors as they were often obtained after the assessment of primary outcome. Future work needs evaluate the properties of the statistic for continuous or time to event outcomes

and formally compare the operating characteristics of this data driven approach to covariate selection to other available methods, particularly in terms of type I error.

Chapter 6

Overall Conclusions

This proposed work offers insight into an important issue that has been overlooked by much the methodological literature in the non-inferiority setting, namely the impact of covariate adjustment. Too often, trialists have taken the lessons learned from superiority and applied them directly to the non-inferiority framework. However, this practice is inherently flawed due to the differential specification of the hypotheses[5], and in fact, for many of the effects found in superiority, just the opposite is true in non-inferiority. In this setting, failing to adjust for covariates results in treatment estimates that are biased towards zero with standard errors that are deflated. This results in some increase in power, but also in a dramatic increase in type I error rates in the non-inferiority setting.[27] This presents a real danger for those trialists who may have unknown covariates or may prefer unadjusted analyses for parsimony or more straightforward interpretation.

In RAMPART, the pre-hospital setting of the intervention made covariate adjusted randomization impossible. Therefore complete randomization and an unadjusted primary analysis were conducted.[40] However, the current work suggests that the treatment effect estimate may be more precise and accurate given a prognostic covariate even if it were unknown to investigators. Results from Aim 1 suggest that it would have been appropriate to adjust for covariates in the analysis even if adjustment at randomization was impossible as this results in unbiased treatment estimates and nominal type I error. Given that no known covariates were available at the start of the trial, an alternate method may have been that employed by Aim 2. This data driven approach would have allowed us to investigate possible covariate effects in the data without knowing their prognostic value in advance and may have resulted in less extreme increases in type I error. Just as it is recommended to conduct both intention to treat (ITT) and per protocol (PP) analyses in non-inferiority trials, trialists may also wish to present both unadjusted and adjusted analyses according to this method when covariates are not known in advance. Thus, this proposed dissertation work (1) quantifies the impact of failing to adjust for covariates in terms of bias, power, and type I error, (2) develops a data driven joint statistic to quantify the impact of adjustment for a covariate given its influence and disparity, and (3) evaluates its potential via practical application.

References

1. Altman, D.G. Comparability of randomized groups. *The Statistician*, 34(1): 125-136, 1985.

2. Altman, D.G. and Dore, C.J. Randomization and baseline comparison in clinical trials. *Lancet*, 335: 149-153, 1990.

3. Beach M.L. and Meier P. Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials*, 10: 161S-175S, 1989.

4. Berger, V. A Novel criterion for selecting covariates. *Drug Information Journal*, 39: 233-241, 2005.

5. Blackwelder, W.C. "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, 3: 345-353, 1982.

6. Canner, P. Choice of covariates in the adjustment of treatment effect. Presented at the Society for Clinical trials Annual Scientific Sessions, San Francisco, CA, 1981.

7. Canner, P. Covariate adjustment of treatment effects in clinical trials. *Controlled Clinical Trials*, 12: 359-366, 1991.

8. Ciolino, J.D. et al. Measuring continuous baseline covariate imbalances in clinical trials data. *Statistical Methods in Medical Research*, 0(0): 1-18, 2011.

9. Ciolino J.D. et al. Covariate imbalance and adjustment for logistic regression analysis of clinical trial data. *Biopharmaceutical Statistics*, 23(6): 1383-1402, 2013.

10. Committee for Proprietary Medicinal Products for Human Use (CHMP). Guideline on the choice of the non-inferiority margin. *Statistics in Medicine*, 25: 1628-1638, 2006.

11. D'Agostino, R. B., Massaro, J. M., and Sullivan, L. M. Non-Inferiority trials: Design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine*, 22: 169-186, 2003.

12. Dunnett, C. W. and Gent, M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometics*, 33,4: 593-602, 1977.

13. Durkalski, V., Silbergleit, R., and Lowenstein, D. Challenges in the design and analysis of non-inferiority trials: A case study. *Clinical Trials*, 0: 1-8, 2001.

14. Fisher, R. A. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33: 503-513, 1926.

15. Ford I., Norrie J., and Ahmadi S. Model inconsistency, illustrated by the cox proportional hazards model. *Statistics in Medicine*, 14: 735-746, 1995.

16. Gail M.H., Tan W.Y., and Piantadosi S. Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75(1): 57-64, 1988.

17. Gail M.H., Wieand S., and Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3): 431-444, 1984.

18. Garrett, A. D. Therapeutic equivalency: fallacies and falsification. *Statistics in Medicine*, 22: 741-762, 2003.

19. Hill, A. B. The clinical trial. *The New England Journal of Medicine*, 247(4): 113-119, 1952.

20. Hilton, J.F. Noninferiority trial designs for odds ratios and risk differences. *Statistics in Medicine*, 29: 982-993, 2010.

21. Hung et al. Some fundamental issues with non-inferiority testing in active controlled trials. *Proceedings of the Annual Meeting of the American Statistical Association*, Aug 5-9, 2001.

22. International Conference on Harmonization. Statistical principles for clinical trials (ICH E9). Food and Drug Administration, DSSH, Febuary 1998.

23. Kahan, B.C. and Morris, T.P. Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine*, 31: 328–340, 2012.

24. Lee Y. and Nelder J.A. Conditional and marginal models: Another view. *Statistics Science*,19: 219-228, 2004.

25. McEntegart, D. J. The pursuit of balance using stratified and dynamic techniques: An overview. *Drug Information Journal*, 37: 293-308, 2003.

26. Neuhaus J.M. and Jewell N.P. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, 80(4): 807-815, 1993.

27. Nicholas K. et al. The impact of covariate adjustment at randomization and analysis for binary outcomes: understanding differences between superiority and noninferiority trials. *Statistics in Medicine*, 34(11): 1834-1840, 2015.

28. Nicholas K., Ramakrishnan, V., and Durkalski, V. Choosing covariates for adjustment in non-inferiority trials based on influence and disparity. Presented at the International Chinese Statistical Association Joint Applied Statistics Symposium. Fort Collins, CO, June 2015.

29. Permutt, T. Adjustment for covariates. *Encyclopedia of Biopharmaceutical Statistics*, 18-21, 2003.

30. Piantadosi, S. Clinical Trials: A Methodological Perspective, 2nd Ed. Wiley Interscience, 2005.

31. Pocock, S.J. et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine*, 21: 2917-2930, 2002.

32. Raab, G. M., Day, S., and Sales, J. How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials*, 21: 330-342, 2000.

33. Robinson, L.D. and Jewell, N.P. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59(2): 227-240, 1991.

34. Romel, J. Therapeutic equivalence investigations: Statistical Considerations. Statistics in Medicine, 17: 1703-1714, 1998.

35. Senn, S. J. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, 8: 467-475, 1989.

36. Senn, S. J. Testing for baseline balance in clinical trials. *Statistics in Medicine*, 13: 1715-1726, 1994.

37. Senn S.J. Conditional and marginal models: Another view – Comments and Rejoinders. *Statistics Science*, 19: 228-238, 2004.

38. Shao, J., Yu, X., and Zhong, B. A theory for testing hypotheses under covariate-adaptive randomization. *Biometika*, 29(2): 347-360, 2010.

39. Silbergleit, R. et al. Intramuscular versus intravenus therapy for pre-hospital status epilecticus. *New England Journal of Medicine*, 366: 591-600, 2012.

40. Silbergleit, R., Lowenstein, D., and Durkalski, V. Study Protocol: A double-blind randomized clinical trial of the efficacy of IM midazolam versus IV lorazepam in the pre-hospital treatment of status epilepticus by paramedics (RAMPART: the Rapid Anticonvulsant Medication Prior to Arrival Trial), Version3. Aug 18, 2010.

41. US Department of Health and Human Sciences, Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research. FDA guidance for industry non-inferiority clinical trials, March 2010.

42. Weins, B. Randomization as a basis for inference in noninferiority trials. *Pharmaceutical Statistics*, 5: 265-271, 2006.

43. Welch, R.D. et al. Intramuscular midazolam versus intravenous lorazepam for the prehospital treatment of status epilepticus in the pediatric population, *Epilepsia*, 56(2): 254-262, 2015.

44. Zhao W. et al. Quantitative comparison of randomization designs in sequential clinical trials based on treatment balance and allocation randomness. *Pharmaceutical Statistics*, 11(1): 39-48, 2012.